2011

Role of Non-Coding RNA NC4 in MLL and CYP33 Mediated Regulation of HOXC8

Jessica Arvindbhai Solanki
Loyola University Chicago

LOYOLA UNIVERSITY CHICAGO

ROLE OF NON-CODING RNA *NC4* IN MLL AND CYP33 MEDIATED REGULATION OF *HOXC8*

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE GRADUATE SCHOOL

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

PROGRAM IN MOLECULAR AND CELLULAR BIOCHEMISTRY

BY

JESSICA A. SOLANKI

CHICAGO, IL

DECEMBER 2011

ACKNOWLEDGEMENTS

I hereby take this opportunity to thank all the people that have influenced in one way or another, my graduate school life, both academically and otherwise. I would like to begin first with the Program in Molecular and Cellular Biochemistry (Division of Cell Biology, Neurobiology and Anatomy) of Loyola University Chicago that allowed me to be a part of their prestigious graduate program, an opportunity invaluable to an International student. I would then like to immensely thank my mentor Dr. Manuel O. Diaz, who accepted me as a student in his laboratory providing me with a second home away from home. He has been amongst the mentors that allow for independent thinking. The one aspect of his personality that I find very rare is his willingness to help and I always found his office door open for all the help I needed. I would like to thank all the past and the present members of his lab that together formed, including me, a very heterogeneous yet a harmonious group willing to help and support each other. In addition to the graduate program I was enrolled in, I received all the help I needed from the Leukemia Research Group at the Cardinal Bernardin Cancer Center. I cannot thank the staff and all the members in the Leukemia Program enough. Lorelei, who is the secretary for the leukemia research group at the cancer center, has always been prompt at extending that helping hand right when I needed it. Thanks to her also for all her efforts in organizing the research events and festivities celebrated by the group.

My heartfelt gratitude towards Dr. Seth Robia (Department of Physiology, Loyola University Chicago) who not only allowed me access to the fluorescence imaging facility located in his laboratory but also guided me through the experimental set up and data analysis pertaining to the second aim proposed in this study that addresses a critical question towards the hypothesis. I could not have come across a more amiable person than him.

I would also like to extend my gratitude towards my dissertation committee that has been critical yet helpful and supportive during some tough times and delays I faced during my dissertation work. They ensured that the project proposed was progressing adequately. I must also thank the graduate program director, Dr. Mary Manteuffel, who is highly supportive of all the students in the Molecular and Cellular Biochemistry program and never fails to guide through the graduate school procedures and ask about personal well-being. I extend my gratitude towards Dr. Simmons, our current graduate program director, who ensures efficient functioning of the program.

Special thanks go to Dr. Richard Schultz for his expertise in enzymology, simulation modeling, protein structure analysis and data interpretation for the experiments from Aim 2 in this dissertation. Thanks also to Dr. Allen Frankfater who provided a lot of insight into the gene expression and regulation aspect of this project.

I would like to extend my gratitude towards Dr. John Bushweller from University of Virginia who agreed to collaborate with our lab with regard to this project. Data obtained from experiments done in collaboration with him pertaining to Aim 2 in this study have helped significantly towards the central hypothesis.

The Division of Molecular and Cellular Biochemistry (MCB program) always gave me the cozy feeling of belonging. The core faculty of MCB comprised of Dr. Manuteuffel, Dr. Collins, Dr. Frankfater, Dr. Simmons and Dr. Schultz are very supportive of their students. The one aspect of the MCB program that I found unique was the way in which the Journal Club was conducted. It allowed for critical analysis of research articles that I believe contributes immensely towards making of a researcher. Apart from the academic community of the program, all the present and past staff members have been like family. I cannot thank Ashiya and Elayne enough for all their help ranging from the filing of necessary forms for graduate school affairs through celebrating birthdays and sometimes socializing with the students. Even after my enrollment at the MCB program, the Department of Cell Biology, Neurobiology and Anatomy (CBNA) has been a second home. The CBNA program staff member Ginny was the first person I came across as being warm and helpful when I joined Loyola University Chicago. Thanks also to Margarita and Judith from the Graduate School office who make the whole system of graduate school studies functioning efficiently. I must also thank Dr. Frederick Weizmann for his support as a Dean through which I was allowed a considerable break from my graduate studies in order to attend to personal affairs back at home.

I would like to thank the Graduate School Office of Loyola University Chicago for supporting me financially through the first two years and of course none of this would have been possible without the NIH grants that support the research in Dr. Diaz's laboratory. I have been a fortunate receiver of the very prestigious Arthur J. Schmidt

Dissertation Fellowship for the 2005-2006 academic years that not only helped me financially but also provided me with an opportunity of Grant Proposal writing.

I must also not forget to thank the Loyola University Health Sciences Library for all their help with my literature search, review and the writing part of this dissertation. The research articles and books that I could not find at the library were always provided to me upon request at the earliest possible via the inter-library loaning system. These have been subtle yet critical steps on my way through graduate studies that ensured an efficient functioning as far as my academic life was concerned.

Just as much fortunate as I have been to be able to pursue my graduate studies, equally blessed I have been to have a family and circle of friends that never failed to show support and love that saw me through both my happy and enduring years. I hope I remain blessed for the years to come.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Co-IP Co-Immunoprecipitations

Ch-IP Chromatin Immunoprecipitations

EMSA Electrophoretic Mobility Shift Assays

NMR Nuclear Magnetic Resonance

HD HomeoDomain containing transcript

CPC Coding Potential Calculator

ABSTRACT

MLL or *Mixed Lineage Leukemia* gene is clinically known for its involvement in genetic translocation with more than 70 different partners identified each giving rise to a highly leukemogenic fusion protein. With a poor prognosis of MLL related leukemia, an investigation into its role during hematopoiesis has been a very active field of research. In order to design strategies to combat the MLL leukemia, it becomes essential to delineate the molecular mechanisms behind the function of MLL wild type protein. Wild type MLL is a transcriptional maintenance protein from the Trithorax Group (TrxG) that resides in complex with other chromatin regulators to maintain the downstream target genes, in particular the *HOX* genes, in their active state of transcription. A nuclear cyclophilin called CYP33 interacts with the third Plant Homeo Domain (PHD) of MLL via its N terminal RNA Recognition Motif (RRM) such that its overexpression switches the function of transcriptional activator protein MLL towards transcriptional repressor. This repression is mediated in part by an enhanced recruitment of histone deacetylase 1 to MLL and a subsequent decrease in the histone H3 acetylation at the MLL target gene promoters, including the *HOXC8* promoter. CYP33 RRM binds Poly A and Poly U RNA in addition to MLL. A regulatory region of MLL target gene *HOXC8* is transcribed into a non-coding transcript, which we call *NC4* in this study. We hypothesize that the regulation of *HOXC8* via MLL may involve *NC4* function such that its expression would

disrupt CYP33 and MLL interaction and consequent repression at the gene promoter. In order to address this, first a more conserved region between *HOXC8* and *HOXC6* was analyzed for transcription into non-coding RNA based on the presence of ESTs and in silico analysis followed by the monitoring of transcription in human and mouse cell lines as well as in mouse embryos. After comparing the expression kinetics of *NC4* and *HOXC8* during in vitro differentiation of mouse embryonic stem (mES) cells, *NC4* transcription was found to precede and exceed that of the *HOXC8* during the mES differentiation implying its early role during differentiation. In order to determine if CYP33 binds a specific RNA sequence, an in vitro selection and enrichment technique was performed that identified a YAAUNY consensus RNA sequence preferentially bound by CYP33. An NMR based collaborative study revealed an overlap between both MLLPHD3 and core AAU RNA sequence binding surfaces on CYP33 RRM that also suggested an exclusive binding by MLLPHD3 and RNA. Furthermore, CYP33 preferred YAAUNY motifs were found to be present at a greater density in the 3' region of *NC4* suggesting its possible role in *HOXC8* regulation by MLL and CYP33. This possibility was first tested using in vitro competition assays based on Forster Resonance Energy Transfer method of studying molecular interactions. YAAUNY containing RNA sequences (both synthetic as well as endogenous *NC4* 3' region) could disrupt the CYP33-MLLPHD3 interaction by 40.66% and 35.78% respectively. Following this, MSA cells that lack endogenous expressions of both *HOXC8* and *NC4* and show the binding of MLL and CYP33 at their promoters, expressed the *HOXC8* gene after ectopic expression of a *NC4* transcript from a transfected plasmid.

CHAPTER ONE

INTRODUCTION

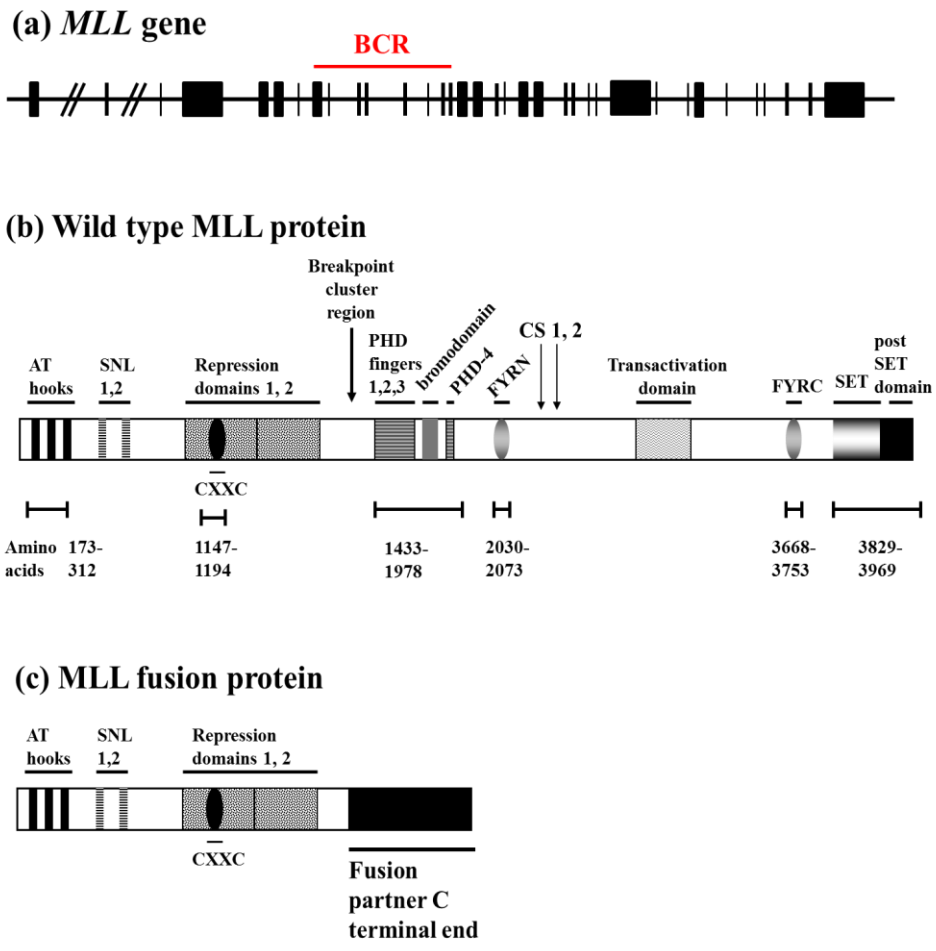
Recurring chromosomal translocations are frequently associated with hematologic malignancies, e.g., the reciprocal translocation t(9;22) forming BCR-ABL fusion results in a chronic myeloid form of leukemia as well as acute lymphoblastic leukemia. Other well-characterized chromosomal rearrangements in translocations in leukemia include t(8;21), t(15;17), inv(16) and translocations involving *MLL* on 11q23 (Zhang and Rowley 2006).

The *MLL* gene is present on human chromosome 11, band q23, encompassing 115.35 Kb of the genomic DNA and coding for a 430 kDa protein [Fig. 1a and b]. Translocation breakpoints are clustered within a 8.3 kb region in the middle of the gene (Ziemin-van der Poel et al. 1991, McCabe et al. 1992). Splicing of fusion transcripts in frame generates a gene fusion product that includes the N terminal domains of MLL and C terminal regions of the partner gene [Fig. 1c]. There have been more than 70 *MLL* translocation partners of which at least 50 have been cloned (Marschalek 2011). The *MLL* translocation partners can be broadly classified as either nuclear proteins involved in the

transcriptional regulation or cytoplasmic proteins that contribute to the dimerization and oligomerization potential of the chimera. Studies undertaken to elucidate the mechanism of leukemogenesis by the above mentioned chimeric proteins point towards a deregulated transcription process that involves protein complexes normally interacting with the partner gene product (Marschalek 2011). This implies that the N terminal-translocated domains of MLL may contribute the target site binding property whereas the partner protein recruits transcription activation complexes. Recent studies involved identification of the roles of MLL fusion specific transcription complexes for more frequently occurring MLL-AF4 and MLL-ENL fusions (Yokoyama et al. 2010, Lin et al. 2010). These studies (Yokoyama et al. 2010. ,Lin et al. 2010) led to an identification of a transcription elongation complex (AEP) containing AF4, AF5q31, Cyclin T1 and CDK-9 that was found bound to the promoter of the MLL regulated target gene Meis1 in both MLL fusion expressing cell lines as well as cells with wild type MLL. MLL fusion oncoproteins, however, recruit the AEP complexes constitutively thereby suggesting both physiological as well as a pathological role of AEP in the MLL target gene regulation (Yokoyama et al.). An MLL fusion mediated effect is the upregulation of *HOXA7*, *HOXA9* and *MEIS1* expression. Upregulation of *Hoxa9* and *Meis1*, independent of the presence of MLL fusion, is sufficient to block myeloid differentiation as well as to promote hematopoietic progenitor cell renewal both of which serve as critical steps in leukemogenesis (Shimamoto et al. 1998, Shah and Sukumar 2010, Kroon et al. 1998, Mohan et al. 2010, Faber et al. 2009).

Figure 1. The MLL gene encodes a 430 kDa multi-domain protein.

(a) The *MLL* gene is 115.35 kb long, located on human chromosome 11 and contains 36 exons shown here as black vertical lines and rectangles. An 8.3 kb Breakpoint Cluster Region (BCR) spans exons 8 through 15 of *MLL* with the break occurring anywhere in the BCR. (b) The *MLL* gene codes for a 430 kDa protein that undergoes post-translational processing, generating 320 kDa N-terminal and 180 kDa C-terminal fragments. The protein is cleaved at two cleavage sites (CS 1 and 2) by a threonine aspartase called the Taspase 1. The two fragments thus generated interact non-covalently via the conserved FYRN and FYRC domains. The various domains identified are mentioned here with their corresponding amino-acid address in the protein. The BCR, in the MLL protein, maps to a region between the repression domains and the PHD finger cassette. (c) In an MLL fusion protein, all the domains of MLL after the repression domains are absent. Shown in the schematic here is a representation of the location of the fusion partner in an MLL fusion protein which is generated as a result of a reciprocal translocation between MLL on chromosome 11 and any one of the 70 different partner genes identified to date.


(a) MLL gene
BCR
(b) Wild type MLL protein
Breakpoint cluster region
PHD fingers 1,2,3
bromodomain
PHD-4
FYRN
CS 1, 2
AT hooks
SNL 1,2
Repression domains 1, 2
Transactivation domain
FYRC
SET
post SET domain
CXXC
Amino acids
173-312
1147-1194
1433-1978
2030-2073
3668-3753
3829-3969
(c) MLL fusion protein
AT hooks
SNL 1,2
Repression domains 1, 2
CXXC
Fusion partner C terminal end

The wild type MLL protein is synthesized as a 430 kDa pro-polypeptide [Fig. 1b] that undergoes a post translational cleavage at two adjacent sites (aa 2666 and aa 2718) in its sequence by a threonine aspartase called Taspase 1 (Hsieh, Cheng, and Korsmeyer 2003). The two peptides (MLL-N and MLL-C) thus generated, re-associate non-covalently via the conserved FYR motifs present in both the peptides thereby conferring stability and nuclear localization to the MLL protein (Hsieh, Cheng, and Korsmeyer 2003). MLL-N contains several motifs and domains: the N terminal AT hook motifs followed by the repression domain that contains the conserved CXXC motif. Following the repression domain in MLL-N is the Plant Homeo Domain (PHD) finger cassette comprised of four conserved zinc fingers and an atypical bromodomain present between the third and the fourth PHD fingers, and finally the FYRN domain. The MLL-C contains an activation domain followed by the FYRC domain and lastly the SET (Su(var)-39, Enhancer of Zeste and Trithorax group) domain. The N-terminal AT hooks bind AT rich, bent and cruciform DNA (Zeleznik-Le, Harden, and Rowley 1994) . The CXXC domain binds unmethylated CpG dinucleotides (Erfurth et al. 2008) and the third PHD finger binds trimethylated H3K4 (Park et al. 2010). Together these domains contribute to the binding of MLL or its fusion proteins to their target loci. In addition, proteins such as LEDGF (Lens Epithelium Derived Growth Factor) that binds nucleosomes and Menin that links MLL to LEDGF are both required for the binding of MLL to its targets (Yokoyama and Cleary 2008). The repression domain binds co-repressor proteins such as Bmi1, HPC2, CtBP and HDAC1 (Xia et al. 2003). The third PHD finger of MLL-N, in

addition to binding to the H3K4me3, binds CYP33 and only recently these interactions have been shown to be exclusive of each other (Park et al. 2010).

The activation domain in MLL-C binds the co-activator proteins CBP (CREB binding protein) (Ernst et al. 2001) and MOF (Males absent On the First) and enhance transcription in a reporter gene assay (Dou et al. 2005). Both CBP and MOF have histone acetyl transferase activity specific for the histones H3/H4 and H4K16 respectively (Rea, Xouri, and Akhtar 2007). The C terminal SET domain exhibits a methyl transferase activity specific for the H3K4 (Milne et al. 2002).

MLL belongs to the trithorax group of maintenance proteins that were initially identified in *Drosophila* and found to be involved in the maintenance of the open state of genes activated for transcription (Yu et al. 1998). Another group of proteins called Polycomb group (PcG) performs the opposing function of maintaining genes in their silent state. MLL functions primarily as a transcriptional maintenance factor that methylates histone H3K4 at the target gene promoter. Taking into account the protein partner profile of MLL, it appears that MLL complex is comprised of both co-activator and co-repressor proteins. Homozygous deletion of *MLL* in mouse causes embryonic lethality by day 10.5 (Yu et al. 1998). The *MLL* null embryos initiate *Hox* gene expression in the appropriate spatial patterns during the first 8 days of development; however, the expression is lost by day 9 indicating a role of *MLL* in the maintenance of *Hox* gene expression as opposed to its initiation (Yu et al. 1998). Furthermore, a *MLL* null mutation rescues a polycomb mutant phenotype indicating its antagonism towards

the silencing function of the polycomb group of genes (Hanson et al. 1999). The *Drosophila* homolog of *MLL* called *Trithorax* or *Trx* is essential during embryogenesis and antagonizes the function of polycomb group of genes during the determination of segmental identity (Cernilogar and Orlando 2005, Ringrose and Paro 2007).

In many aspects of structure as well as function, MLL resembles the *Drosophila* protein trithorax or trx. Trithorax shares the sub nuclear localization signals, the PHD finger region, post translational cleavage by Taspase 1, FYRN and FYRC domains as well as the C terminal SET domain with MLL, however, it lacks the N terminal AT hooks and the CXXC domains. With its cohort of proteins such as dSbf1, dCBP, Spt5, Spt6 and FACT (for Facilitates Chromatin Transcription), trx regulates the expression of Hsp70 at the level of transcription elongation (Smith et al. 2004). This feature of trx mediated gene regulation occurs via binding of the trx complex to the Hsp70 promoter. Apart from binding to the target gene promoter, trx and certain Polycomb group (PcG) proteins bind to the intergenic regions of *Drosophila Hox* genes that contain Polycomb/Trithorax regulatory elements (PRE/TREs) (Schuettengruber et al. 2009, Schwartz et al. 2010). It is believed that the recruitment of TrxG and PcG proteins occurs via sequence specific DNA binding proteins (Ringrose and Paro 2007). These have been identified as Pho (Pleiohomeotic), Dsp1 (Dorsal switch protein 1), GAF (GAGA factor), Zeste, Grh (Grainyhead) and Sp1 (Specificity protein 1) in *Drosophila* species. In mammals, however, the identification of TRE/PRE DNA binding proteins has been limited to YY1 (Brown et al. 2003, Atchison et al. 2003), which is a Pho homologue and a *Drosophila* GAF homologue called Krox (Matharu et al. 2010). From numerous studies in

Drosophila, it is becoming increasingly clear that the maintenance proteins come into play early during development even before the initiation factors such as the maternal effect and segmentation gene products disappear after preparing the chromatin for specific genetic programs in an embryonic domain specific manner (Ringrose and Paro 2007, Cernilogar and Orlando 2005). All the gene regulatory circuits require the role of cis regulatory elements in specifying the time, tissue, cell type and the extent of gene expression. These cis elements in clustered genes are present in the intergenic region and in *Drosophila* these contain PRE/TREs (Schuettengruber et al. 2007). Mammalian equivalents of *Drosophila* TRE/PRE have not been fully characterized yet. Studies by Erfurth et al. (Erfurth et al. 2008) suggest a role for MLL in the protection of CpG islands from methylation at the target gene promoters. Furthermore, MLL interaction with histone acetylases such as CBP and MOF leads respectively to H3 and H4 acetylation which serves as a critical layer of the transcriptional activation process. Correlating with the co-occupancy of TrxG and PcG proteins at the TRE/PRE in *Drosophila* (Orlando et al. 1998, Ringrose and Paro 2007) is the presence of bivalent chromatin domains marked by both the activation mark H3K4me3 and the repressive histone modification H3K27me3 in mouse embryonic stem cells (Bernstein et al. 2006). It was recently elucidated for the early lineage cells such as the Hematopoietic stem cell and Progenitor cells HSCs/HPCs that the bivalent chromatin is resolved into either transcriptionally active genes or silent genes with each category marked by a specific combination of histone modifications (Cui et al. 2009).

Both MLL and Trithorax are capable of binding via their PHD finger region to CYP33, a protein that shifts their function from transcriptional activation to repression (Anderson et al. 2002, Fair et al. 2001). In humans, further studies showed that CYP33 mediates repression of *HOXC8*, a MLL target gene, by enhancing the recruitment of histone deacetylases HDAC1 and HDAC2 to the repression domain of MLL (Xia et al. 2003).

MLL regulates *Homeobox (HOX)* genes during both embryonic development and hematopoiesis. *Hox* genes exist in clusters in *Drosophila* as well as in vertebrates. In mammals for example, 39 *Hox* genes are distributed over four separate clusters on different chromosomes. In *Drosophila*, two separate *Hox* clusters exist namely, the *Antennapedia* complex (*ANTP-C*) and the *Bithorax* complex (*BX-C*), both present on chromosome 3R (right arm). *Hox* genes (in both *Drosophila* and mammals) carry regulatory elements that are involved in ensuring correct spatial and temporal expression patterns of their cognate genes during development and differentiation. More particularly, the temporal expression of *Hox* genes throughout development follows an order related to their placement in the cluster with the 3' genes being expressed earlier than the more 5' genes (Deschamps et al. 1999). A similar regulatory process holds true for the *Drosophila BX-C* genes. These genes are under the regulation of elements present in the intergenic regions that specify the spatial and temporal expressions in specific cell types during various stages of development. In *Drosophila*, the cis-regulatory elements involved in body patterning have been identified as 'initiator elements', 'maintenance elements' and tissue specific/cell type specific' enhancer elements based on their function

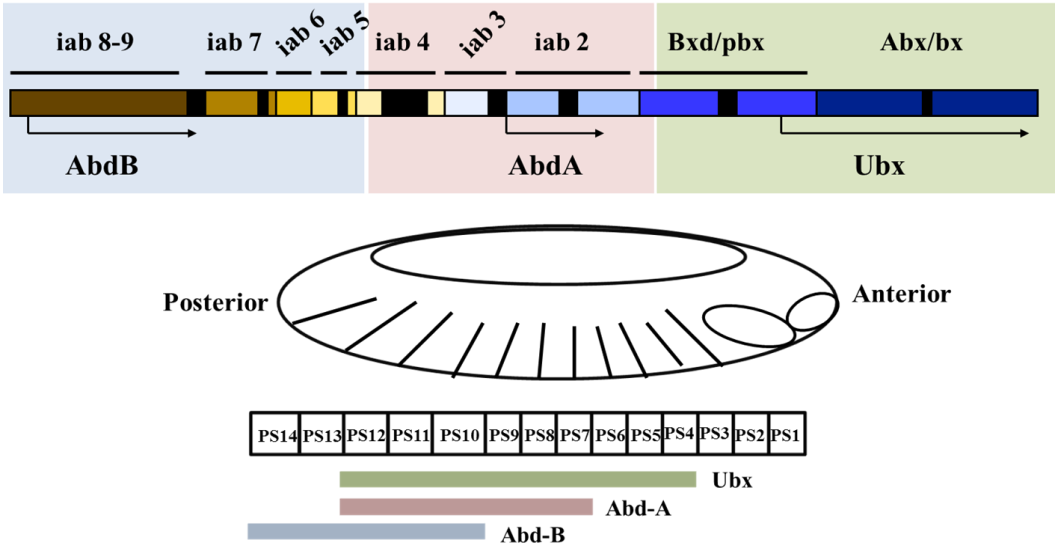
during the course of development (Deutsch, 2010). The regulatory elements involved are called the infra-abdominal enhancers (iab), promoter targeting sequences (PTS) and the expression-restricting boundary elements MCP and the front-abdominal regions (fabs) [Fig. 2]. In the fly, iabs are involved in providing the segmental address to coding *Hox* genes by directly interacting with their associated promoters. The fabs function in restricting the promoter-enhancer interactions in the segments that must lack the expression of a given *Hox* gene (Zhou et al. 1999) whereas the PTS overcome the restriction imposed by boundary elements for certain promoters (Zhou,J. 1999). With mutational analysis, three boundary elements have been defined namely, Fab 7, Fab 8 and Mcp (Maeda and Karch 2006, Barges et al. 2000, Gyurkovics et al. 1990, Mihaly et al. 1998, Hagstrom, Muller, and Schedl 1996, Zhou et al. 1996, 3195-3201, Gruzdeva et al. 2005, 3682-3689). In particular, mutations in the iab elements correspond to homeotic transformations of the embryonic segments that they regulate. Corresponding to an iab mutation, the *Hox* gene that it regulates is misexpressed in the iab controlled domain (Maeda and Karch 2006).

In addition to these elements acting in cis to regulate gene expression, these are themselves transcribed and the occurrence of these non-coding transcripts (ncRNA) in *Drosophila* have been known for more than two decades (Lipshitz, Peattie, and Hogness 1987, Bae et al. 2002, Petruk et al. 2007, Rank, Prestel, and Paro 2002, Cumberledge, Zaratzian, and Sakonju 1990, Tupy et al. 2005). Many of these studies done to unravel the function of ncRNA found them expressed in a regulated manner during development (Cumberledge, Zaratzian, and Sakonju 1990, Bae et al. 2002, Akbari et al. 2006, Rank,

Prestel, and Paro 2002). ncRNA in *Drosophila* arise from two classes of enhancers namely the early embryonic enhancers and the late larval enhancers (Lempradl and Ringrose 2008). Comparison of the ncRNA expression to the coding transcripts reveals that its expression precedes that of the coding gene; however, the expression from both is limited to the same embryonic domains (Bae et al. 2002). This led to the hypothesis that enhancers carrying the regulatory elements TRE/PREs are expressed first to define the segmental identity in a developing embryo following which, the expression of coding transcripts is maintained in the corresponding segments by the TrxG of proteins (Bae et al. 2002). Until recently, studies were lacking that could explain how ncRNA may be involved in the regulation of their associated coding genes. Studies in *Drosophila* reveal that even though the expression of *bxd* ncRNAs from the *bxd* element as well as the *Ubx* (*Ultrabithorax*) expression both occur in halteres and 3rd leg discs, at a single cell resolution; however, the expression of the two is never seen occuring in the same cell nucleus implying an inhibitory role of *bxd* ncRNA towards *Ubx* expression (Petruk et al. 2006). Another study (Sanchez-Elsner et al. 2006) yielded contrasting results demonstrating a role for three different *bxd* ncRNA in the regulation of *Ubx* via recruitment of the TrxG protein called Ash1. This group showed recruitment of Ash1 to the TRE template via the binding of its SET domain to the ectopically expressed *TRE* RNA thereby activating the neighboring *Ubx* gene in tissues that lacked endogenous *TRE* transcript expression (Sanchez-Elsner et al. 2006). In the past three years, more studies have identified regulatory RNAs, in addition to the protein and DNA, as a third important component involved in control of gene expression at the chromatin level. In the context

of *HOX* genes in humans, two separate long ncRNA (>200 nucleotides) from *HOX* loci were shown to be involved in the regulation of coding *HOX* genes in trans after binding to chromatin modifying proteins (Rinn et al. 2007, Wang et al. 2011). For many other genetic loci, studies have shown a correlation between coding gene expression and that of non-coding transcripts (Zhang et al. 2009, Dinger et al. 2008, Mercer et al. 2008). It has been observed for clustered genes such as the *Hox* genes, *β-globin* genes, *myosin heavy chain* genes and the *interleukin* genes, that the intergenic regulatory elements are transcribed into non-coding transcripts (Plant, Routledge, and Proudfoot 2001, Rogan, Cousins, and Staynov 1999, Haddad et al. 2008, Ho et al. 2009, Rank, Prestel, and Paro 2002). For mammalian *Hox* genes, the relationship between the expressions of coding *Hox* genes as well as the enhancer-transcribed ncRNAs during development is not well studied.

Figure 2: Drosophila *Bithorax* Complex: *BX-C* is one of the two *Homeotic* gene complexes in *Drosophila* that is comprised of three protein coding genes, the expressions of which are controlled in the required domains by cis regulatory elements called *iabs* present flanking the coding gene (shown in red are *iabs* that support *Abd B* expression, *iabs* in black for *Abd A* and iabs in yellow for *Ubx* expressions). The *iabs* function as enhancers of the coding genes whereas the *Fabs* (rectangles) function as insulators that restrict the enhancer-promoter interactions in the wrong domains. The coding genes are depicted as black arrows. Shown in the given schematic are also expression boundaries of the coding genes in different parasegments of a developing *Drosophila* embryo (figures adopted from the references mentioned). *Drosophila* embryo is shown with the parasegments 10, 11 and 12 expressing all three *Hox* genes.



As mentioned earlier, MLL regulates a subset of *Hox* genes in vertebrates during development. *HOXC8*, in particular, has been shown to be a direct target of MLL by both MLL deletion analysis (Yu et al. 1998, Hanson et al. 1999) and by chromatin immunoprecipitation (ChIP) studies that looked into the binding of MLL to the *HOXC8* gene promoter (Milne et al. 2002). *HOXC8* belongs to the *HOXC* gene cluster present on human chromosome 12 and mouse chromosome 1. The expression pattern for *Hoxc8* has been well characterized during mouse embryonic development (Kwon et al. 2005, Deschamps et al. 1999) including some of the cis-regulatory elements that are required for its expression (Bieberich et al. 1990, Shashikant and Ruddle 1996, Anand et al. 2003, Bradshaw et al. 1996, Shashikant et al. 1995, Juan and Ruddle 2003). Like many other *Hox* genes, *Hoxc8* expression during mouse embryogenesis occurs in two phases i.e. an early initiation phase and a late maintenance phase (Deschamps et al. 1999). Two separate regulatory elements have been identified for the two phases; an early phase of expression under the control of an early enhancer located 5' to the *Hoxc8* gene promoter (Shashikant et al. 1995, Shashikant and Ruddle 1996) and a late phase of expression under the regulation of a broad, roughly defined, 8 kb region present 3' to the *Hoxc8* gene (Bradshaw et al. 1996). The maintenance phase of *Hoxc8* in mouse refers to determination of its anterior boundary of expression in the developing embryo and stabilization of its expression in specific cell types within the domains it controls. Since MLL is also required for the maintenance of *Hoxc8* during mouse embryogenesis (Yu et al. 1998, Hanson et al. 1999), it is plausible that the late phase maintenance element present at the 3' region of *Hoxc8* is under regulation of the MLL complex. The *HOXC8*-

HOXC6 intergenic region has other features that suggest its role during development or differentiation. Two minor and a major DNaseI hypersensitive sites that represent protein lodging regions on the locus, have been mapped in this region of which the minor sites overlap with the *Hoxc8* 3' regulatory element (*Hox8* 3' RR in Fig.5). Three regulatory elements have also been identified in this region which, in a transgenic reporter gene assay, potentiates transcription from the *Hoxc8* promoter in mouse embryonic fibroblasts (Milne et al. 2002). Furthermore, a bivalent chromatin domain was found to be present between 7.6 kb – 10.6 kb relative to the *Hoxc8* transcription start site TSS in mouse embryonic stem cells (Bernstein et al. 2006) suggesting its possible regulatory role during development.

As mentioned before, clustered genes express ncRNAs from some of their regulatory elements. Likewise, search for Expressed Sequence Tags (ESTs) revealed the presence of five major ESTs in the 16 kb long intergenic region between *Hoxc8* and its 3' neighbor *Hoxc6* [Fig. 5]. It remains to be tested if these non-coding transcripts have any role in the regulation of the *Hoxc8* gene during development or cellular differentiation. This possibility becomes particularly attractive in the context of CYP33 mediated switch of MLL function from activator to repressor (Anderson et al. 2002, Fair et al. 2001). CYP33 is a nuclear cyclophilin with an N terminal RNA Recognition Motif (RRM), a conserved spacer and a C terminal cyclophilin domain. CYP33 binds MLL via its RRM, however the same domain is also involved in binding to RNA polynucleotides (Mi et al. 1996). It can thus be hypothesized that CYP33 couples silencing mechanisms to the MLL complex and its binding to RNA may sequester it from the complex thereby relieving

repression from the MLL target gene. In order to elucidate whether ncRNA from the *HOXC8-HOXC6* intergenic region have a function in the MLL mediated regulation of *HOXC8*, the following questions were addressed using molecular and biochemical approaches in the current study:

1) Whether the non-coding transcripts mapped in the *HOXC8- HOXC6* spacer region are conserved in mammals and if they truly lack the potential to be translated into proteins. It also becomes essential to determine if their expression is regulated during development and differentiation.

2) Whether CYP33 has a preference for binding to a specific RNA sequence and whether it can bind the endogenous non-coding transcripts under study. Since the RRM of CYP33 can bind both RNA and the MLL PHD finger 3 (MLL PHD3), we tested whether there is a competition between RNA and MLL PHD3 for binding to CYP33.

3) Finally, whether ectopic expression of ncRNA from the *Hoxc8* 3' RR can trans-induce expression of *HOXC8* in a suitable cell line that does not endogenously express *HOXC8*.

Addressing these questions will test the possible regulatory role of non-coding transcripts in modulating the expression of a coding gene like *HOXC8* via the MLL and CYP33 functional interaction.

CHAPTER TWO

Review of relevant literature

MLL translocations and Leukemia

The *MLL* gene located on human chromosome 11 (band q23) fuses with other genes as a consequence of chromosome translocations due to DNA breakage in the 8.3 kb breakpoint cluster region (BCR) [Fig. 1a] caused upon exposure of cells to DNA Topoisomerase II inhibitors, often used in chemotherapy for patients with cancer (Harper and Aplan 2008). The MLL breakpoint cluster region spans introns 8 through 15 including the exons in between. More than 70 different translocation partners have been identified for *MLL* to date (Marschalek 2011). All the MLL fusions retain the N terminal 1/3rd of the protein that mostly provides the domains for binding to the downstream target genes [Fig. 1c]. The partner gene on the other hand contributes an unregulated transcriptional advantage either in the form of dimerization potential or binding of specific protein complexes. When compared, the MLL fusion partners can be classified into dimerization inducing (via leucine zippers and α-helical coiled coil domains) cytoplasmic proteins or nuclear partners that recruit various transcriptional activation

protein complexes (Daser and Rabbitts 2005). Most of the MLL fusion induced leukemias are characterized by an upregulation of *HOXA7*, *HOXA9* and *MEIS1* expression. *HOXA10* and *MEIS1* when overexpressed in the absence of an MLL-fusion gene can also induce leukemia development in mouse models (Ayton and Cleary 2003).

It has also been demonstrated that the immortalization potential is compromised upon down regulation of both *Hoxa7* and *Hoxa9*, indicating that the transformation of myeloid progenitors by MLL oncoproteins is dependent on the overexpression of *Hoxa7* and *Hoxa9* (Ayton and Cleary 2003).

Wild Type MLL protein, its protein interaction partners and their known functions:

The wild type MLL protein [Fig. 1b] is post-translationally cleaved into a 320 kDa N-terminal region (MLL-N) and a 180 kDa C terminal region (MLL-C) by threonine aspartate protease called Taspase 1 (Hsieh, Cheng, and Korsmeyer 2003). WT MLL protein interactions have been studied by either using its individual domains or by biochemically purifying MLL-associated protein complexes from different cell types [Table 1]. The MLL-N terminus contains the AT hooks, sub nuclear localization signals, two repression domains and the Plant Homeo Domain (PHD) finger cassette. The AT hooks as well as the post-SET domain (from the C terminus) binds the INI1 a chromatin protein component of the SWI/SNF nucleosome remodeling complex and SBF1 pseudophosphatase. INI1 functions in chromatin remodeling followed by transcriptional activation or repression (Rozenblatt-Rosen et al. 1998). The three AT hooks form a domain that binds AT nucleotide tracts and bent or cruciform DNA (Zeleznik-Le,

Harden, and Rowley 1994) . The sequences containing the AT hooks bind to proteins such as the GADD34 (Adler et al. 1999), SET (Adler et al. 1997) and protein phosphatase 1 (PP1) (Wu et al. 2002). A region N-terminal to the AT hooks binds Menin, a tumor suppressor gene product (Yokoyama et al. 2004). Proteins such as cMyb (Jin et al. 2010) and LEDGF (Lens Epithelial Derived Growth Factor) (Yokoyama and Cleary 2008) bind MLL via Menin. Menin has proven to be a critical protein interaction partner required for the binding of MLL as well as MLL fusions to its target genes (Yokoyama et al. 2005). Placed next to the AT hooks are four sub-nuclear localization signals (SNLs) that mediate the sub nuclear punctate distribution of the processed MLL protein within the nucleus (Yano et al. 1997). A reporter assay based study revealed presence of a repression domain (Zeleznik-Le and Harden et al. 1994) after the SNLs that also contains a highly conserved CXXC motif with preference for binding to unmethylated CpG DNA and protecting it from DNA methylation mediated repression mechanisms (Erfurth et al. 2008). Repression domains 1 and 2 bind to co-repressor proteins such as the HPC2 (Human Polycomb protein 2), Bmi1 (B lymphoma Mo-MLV insertion region), CtBP (C-terminal binding protein) as well as HDAC1 (Histone Deacetylase 1) and HDAC2 (Xia et al. 2003) in co-immunoprecipitation assays. Recently, PAF1, a component of the Polymerase Associated Factor complex (PAFc) was shown to interact with the sequences flanking the CXXC motif. The role of the PAF complex is to link MLL to RNA polymerase II and stimulate histone modifications (Muntean et al. 2010). Next to the repression domains are four highly conserved Plant Homeo Domain (PHD) fingers

forming a cassette with an atypical bromodomain placed in between the third and the fourth PHD finger.

Table 1: MLL domains and the corresponding interaction partners. Most of the results summarized here are from studies undertaken using individual MLL domains and their characterization in terms of molecular interactions as well as the effect such interactions have on the function of MLL as a transcriptional regulator. All the interactions after the first 1400 amino acids of the wild type MLL are lost in the MLL-fusions. Most studies undertook an approach to demonstrate either direct or indirect protein-protein interactions in vitro and in cell lines respectively. Co-binding of MLL and interaction partners to chromatin has been widely studied using Ch-IP.

MLL domain (amino acids)	Partner domain (amino acids)	Cell system/method	Functional significance of the interaction	Ref.
AT hooks; 142-400	Cruciform dsDNA	EMSA	MLL binds DNA structures	Zeleznik-Le et. al., 1994
1- 1406 1799-1802	Menin Kelch domain of HCF1	Co-IP/ K562 nuclear extracts. Domain mapping/ 293 cells.	Menin binds Y-, branched- and 4- way DNA Function of HCF-1 not known.	Yokoyama et. al., 2004 Yokoyama et. al., 2008
112-153	LEDGF IBD (HIV-1 integrase binding domain)	Co-IP/293T cells Ch-IP/293T cells LEDGF binding site mutation/mouse bone marrow colony formation. ChIP/293T cells for requirement of Menin for interaction of MLL with LEDGF	LEDGF binds acetylated H3 and H4 facilitating binding of protein cargo close to transcription start site	Yokoyama et. al., 2008
RD1; 1147-1203	Palindromic dodecameric DNA with single unmethylated CpG	NMR/direct interaction, mutational studies to identify critical CXXC residues	MLL protects CpG from methylation by binding to unmethylated CpG	Cierpicki et. al., 2010
CXXC containing 1147-1244	HexaCpG ds DNA	EMSA	Protection of unmethylated CpG from methylation	Birke et. al., 2002
RD1: 1101-1250 RD2: 1251-1400	HDAC1, CtBP, HPC2 and Bmi-1 Bmi-1 only	Co-IP/293T cells Reporter gene assays demonstrating repression/293T cells	All proteins cause repression by MLL	Xia et. al., 2003
PHD1-3 bromo domain cassette (1568-1767)	H3K4Me3 /2 and CYP33 (via RRM)	X ray crystallography ChIP/ 293T cells	Tethering of MLL to target chromatin	Wang et. al., 2010

MLL domain (amino acids)	Partner domain (amino acids)	Cell system/method	Functional significance of the interaction	Ref.
CXXC domain containing 1116-1397 fragment	PAF complex containing CDC73, PAF1, LEO1, CTR9 and WDR61	Co-IP/ 293T cell line Ch-IP/ 293T cell line Mouse bone marrow progenitor transformation assay by MLL-PAF complex interaction	PAF complex stimulates transcriptional activity of wild type MLL and MLL fusions	Muntean et. al., 2010
AT hooks, CXXC subdomain and PHD cassette in combination with PAF complex	In vivo *HOXA9* promoter binding	Co-IP/ mouse embryonic fibroblasts (with WT or mutated CXXC and PHD cassette along with PAF complex	MLL binding to target site observed: 1) not with AT hooks alone, 2) partially with AT hooks and CXXC, 3) with AT hooks, CXXC and PHD cassette. Disruption of interaction inhibited *HOXA9* transcription	Milne et. al., 2010
MLLPHD 1-4: 1392-2000	RRM of CYP33	Yeast two hybrid, GST pull downs, Co-IP and co-localizations/293T cand HeLa cell line	Modulation of MLL function towards gene repression	Fair et. al., 2001
M5 fragment of MLL: 3100-3300	MOF: 1-235	Purification of FLAG-WDR5 complex, Co-IP/ 293T cells, Domain mapping/ GST tagged fragments in vitro binding , MOF based HAT assays	H4K16 acetylase activity by MOF H3K4 trimethylation by MLL complex	Dou et. al., 2005
MLL unknown region	CGBP unknown region	Identification of FLAG-CGBP bound HMTase (MLL) activity followed by binding of both proteins to target by Ch-IP	siRNA mediated downregulation of CGBP inhibits MLL target gene expression	Ansari et. al., 2008
2829-2883	CBP; 581-687	GST tagged fragments for direct interaction	Contributes H3 and H4 acetylation function to MLL complex	Ernst et. al., 2001

MLL domain (amino acids)	Partner domain (amino acids)	Cell system/method	Functional significance of the interaction	Ref.
MLL-C	Ash2L, WDR5 and RbBP5	In vitro reconstitution/ Sf9 cell line expressed proteins siRNA based studies in 293 cell line for effect on H3K4 methylation and target gene expression	Core complex needed for H3K4trimethylation. siRNA mediated depletion of individual proteins does not affect MLL-C binding	Dou et. al., 2006 Patel et. al., 2009
MLL-C; 3745-3969 Trx-C; 3540-3969	INI-1; 118-315 SNR1; 168-302	Yeast 2 hybrid for interaction and domain mapping. In vivo binding using radiolabeled SNR1/INI1 proteins and COS cell expressed Trx and MLL respectively	Endogenous Trx and epitope tagged SNR1 co-localize on Drosophila polytene chromosomes. INI1/SNR1 may contribute nucleosome remodeling activity to MLL complex	Rozenblatt-Rosen et. al., 1998

The third of the four PHD fingers (MLLPHD3) binds the nuclear cyclophilin protein called CYP33 (Anderson et al. 2002, Fair et. al., 2001) as well as the H3K4me3 modification (Park et. al., 2010; Wang et. al., 2010). These two interactions with the PHD3 of MLL have been shown to be mutually exclusive (Park et al. 2010). Overexpression of CYP33 to the PHD3 leads to an increased recruitment of histone deacetylase 1 (HDAC1) to the MLL repression domain thereby switching the MLL function to gene repression. MLL fusions lack the PHD3 and hence cannot interact with CYP33. The lack of CYP33 mediated repression at MLL-fusion bound target genes may be one factor that contributes towards leukemogenesis since re-introduction of the PHD finger cassette into MLL fusions such as the MLL-ENL and MLL-AF9 attenuates their transformation capacity (Chen et al. 2008, Muntean et.al. 2008). This clearly indicates that despite the major role of MLL as a maintenance factor for gene expression, proper gene regulation requires an involvement of repressive proteins such as histone deacetylases, the lack of which may lead to unregulated transcription. Further studies are required to study the context dependent co-repressor functions of the MLL complex.

MLL-C is a shorter peptide (180 kDa) and contains an activation domain that is known to bind CBP, a histone acetyltransferase (Ernst et. al., 2001), MOF (Males absent on the First) another histone acetyl transferase specific for H4K16 (Dou et al. 2005, Rea, Xouri, and Akhtar 2007) as well as the SET (Su(var) 39, Enhancer of Zeste and Trithorax) domain that has a H3K4 specific methyl transferase activity (Milne et al. 2002). Association of MOF with MLL is required for an optimal *Hoxa7* and *Hoxa9* transcription (Dou et al. 2005, Milne et al. 2002). Table 1 provides further details on the

domain mapping studies done for both protein interaction partners as well as the functional significance of each interaction. Taken together, it can be hypothesized that the multidomain MLL, in addition to its intrinsic histone methyltransferase activity, serves as a scaffold for various chromatin regulating factors to assemble at the target gene locus in order to orchestrate transcription regulation. In addition to the protein interactions mentioned above, multiple protein containing MLL complexes have been purified from various cell lines like HeLa, K562 and U937 by two groups (Nakamura et al. 2002, Dou et al. 2005). More importantly, compared to the individual domain studies done, these studies have revealed a stable binding of the proteins that depict an active transcription process e.g. the TATA box element binding protein, TBP associated factors, nucleosome remodeling complexes, DNA and RNA helicases, snRNPs (small nucleolar Ribo nucleo protein), proteins involved in mRNA polyadenylation and splicing, WD40 domain containing proteins that mediate protein-protein interactions, histone acetyl transferases and spliceosomal proteins (Nakamura et al. 2002, Dou et al. 2005).

Even though the RNA polymerase II does not purify in any of the complexes characterized above, some studies have reported co-binding of MLL and RNA polymerase II at transcriptionally active loci (Mohan et al. 2010, Hess 2004, Milne et al. 2005, Muntean et al. 2010); whereas others have found RNA polymerase II co-localizing with the H3K4me3 modification that typically occurs at the transcription initiation site of MLL target genes (Wang et al. 2009, Ansari et al. 2009). All the evidence provided above suggests an involvement of MLL in the transcriptional regulation of its target genes.

MLL has an essential function during definitive hematopoiesis (Ernst et al. 2004). It is required for the maintenance of gene expression from the *Hoxa* and *Hoxb* clusters during murine definitive hematopoiesis (Ernst et al. 2004). *MLL* null mouse embryonic stem cells fail to differentiate into hematopoietic colonies, which is rescued by the ectopic expression of *Hoxb3*, *Hoxb4*, *Hoxa9* and *Hoxa10* indicating an MLL dependent role of these genes in hematopoiesis (Argiropoulos and Humphries 2007). In general, certain *Hox* genes are maintained at the required levels of expression in early hematopoietic progenitors including the stem cells. This expression is attenuated during the course of differentiation (Slany 2009). Overexpression of select *Hox* genes in hematopoietic progenitors, deregulation via *MLL* translocations or other genetic abnormalities all can lead to a block in the differentiation pathway thereby serving as initial events in leukemogenesis. As mentioned earlier, the maintenance of the required *Hox* gene expression by MLL both during development and hematopoiesis occurs at the level of chromatin regulation which includes DNA binding, DNA methylation, histone modifications, nucleosome remodeling, transcription initiation and elongation.

Cellular Memory Modules: Lessons from Drosophila

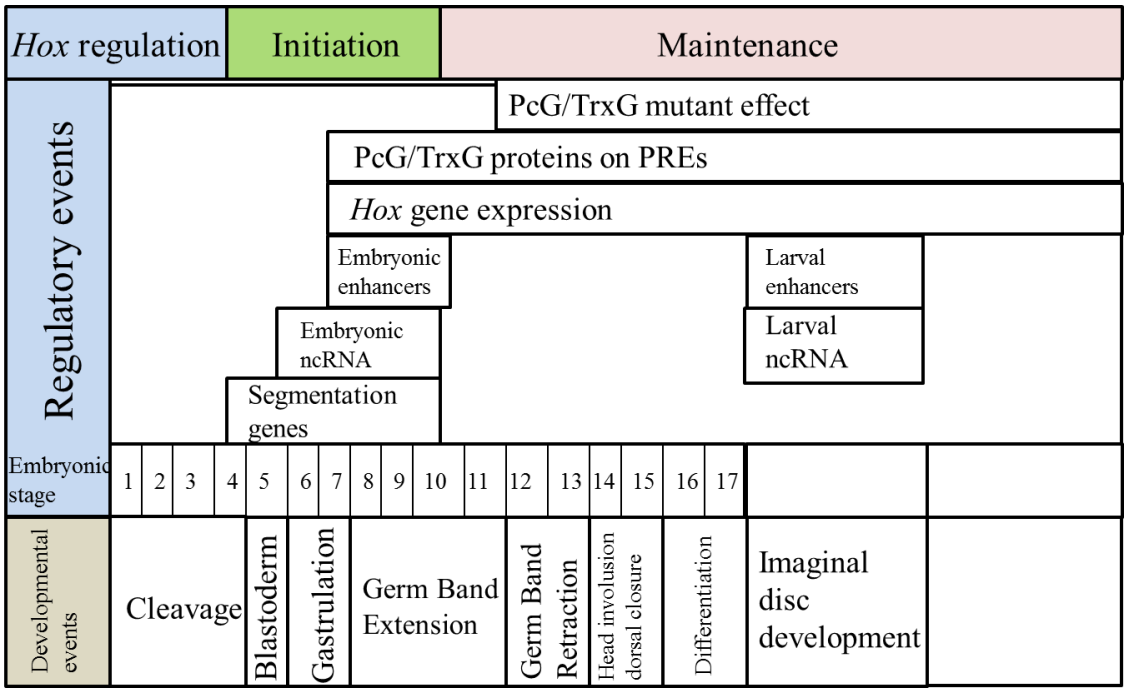
Early lineage cells such as the embryonic and adult stem cells display gene expression profiles characteristic of ongoing transcription required for the self-renewal, proliferation and cell cycle regulation (Boyer et al. 2006, Loh et al. 2006). When cells undergo differentiation, genes for self-renewal are turned off whereas lineage specific genes are turned on. It has been observed in mouse embryonic stem cells that many developmental

regulator genes are characterized by the presence of bivalent histone modifications characteristic of both the active transcriptional state of expression (H3K4me3) and the silenced state (H3K27me3) , (Bernstein et al. 2006). This dual modification represents a poised chromatin state that is receptive to signals for differentiation and eventually resolves into either the active or silent gene expression states (Cui et al. 2009, Mendenhall and Bernstein 2008). One study monitored the changes in gene expression profile and histone modifications that occur when hematopoietic stem/progenitor cells are induced to differentiate into erythrocyte precursors. This model of stem cell differentiation revealed histone modification signatures capable of distinctively identifying enhancers from the promoters that have been prompted for transcription (Cui et al. 2009). Neither of the studies above directly addressed what role TrxG or PcG proteins play in the process of differentiation. However, it is known that H3K4 is trimethylated by SET1, MLL or trithorax whereas the H3K27 is trimethylated by the Polycomb proteins of the PRC2 complex (Schuettengruber et al. 2007, Ringrose and Paro 2007). The TRE (Trithorax Response Elements) and PRE (Polycomb Response Element) have not been fully characterized in mammals yet (Ringrose and Paro 2007). TRE/PREs serve as a molecular address for sequence specific DNA binding proteins that primarily function as adaptors for the TrxG and PcG proteins to bind to the locus (Ringrose and Paro 2007). Studies from *Drosophila* have provided answers to many questions regarding the function of TrxG and PcG proteins during the course of development. Cellular Memory Modules (CMM) contain TRE/PRE that ensure the maintenance of *Hox* gene expression patterns through differentiation. Only recently, much interest has led to

studies that elucidated the role of non-coding transcripts in the regulation of *Hox* genes during *Drosophila* development. Occurrence of non-coding transcripts in *Drosophila* embryos in a regulated manner has been noticed for more than two decades and had remained an enigma until the recent past (Ho et al. 2009, Rank, Prestel, and Paro 2002, Bae et al. 2002).

The enhancers from the *Drosophila Bithorax* complex (BX-C) are transcribed twice during development in a regulated manner [Fig. 3]. The first phase coincides with early embryogenesis during which the embryonic enhancers are transcribed into ncRNA followed by the larval enhancers that are active during a later larval phase (Lempradl and Ringrose 2008). In addition to being transcribed earlier than the coding genes, these transcripts are expressed in the same segments as the coding genes. Deletion of enhancers coding for these transcripts during early embryogenesis leads to homeotic transformations akin to the transformations caused by deletion of *Hox* genes (Lempradl and Ringrose 2008, Bae et al. 2002).

Figure 3: Summary of events occurring during *Drosophila* embryogenesis that depict transcription of *Hox* genes and their control by TrxG/PcG as well as associated enhancers. Shown in this figure adapted from (Lempradl and Ringrose 2008) are the distinct initiation and maintenance phases identified in *Drosophila* embryogenesis that are characterized by the regulated expressions of the *Hox* genes and ncRNA from enhancers. Even though the maintenance proteins are required mainly during the maintenance phase of the *Hox* genes, their binding to their response elements (PRE) occurs much earlier as shown.



These observations, when taken together, suggest a role for early expression of non-coding transcripts in the marking of an open chromatin domain for coding gene expression to follow (Lempradl and Ringrose 2008). This hypothesis adds a new dimension to the function of cellular memory modules, which may rely on the transcription from enhancer elements to regulate the transcription from their associated promoters (Rank, Prestel, and Paro 2002). Some mechanistic aspects of TrxG and PcG mediated maintenance of gene expressions are common between *Drosophila* and mammals such as the protein complexes representing the two groups, the occurrence of *homeobox* genes in clusters as well as the presence of intergenic cis-regulatory elements (Deschamps et al. 1999). For many clustered genes in mammals, regulatory elements have been found to be transcribed into non-coding RNAs (ncRNAs) that are expressed in a tissue specific manner correlating with the expression of coding genes (Rogan, Cousins, and Staynov 1999, Mercer et al. 2008, Xiang et al. 2006, Gribnau et al. 2000, this study). Whether these are regulatory RNAs equivalent to the ncRNAs found in *Drosophila* remains an attractive yet unaddressed question.

Role of Non-coding Intergenic Transcripts

The human genome contains approximately 30,000 protein coding genes (Venter et al. 2001). The proteome, however, achieves its further complexity by virtue of alternative splicing of the mRNA (Graveley 2001). Considering the entire length of the human genome, the transcriptional output from it far exceeds the number of proteins known to be encoded by it (Mattick 2001). About 98% of the transcription from the

human genome lacks protein coding potential thus representing a by far underestimated pool of cellular ncRNA. The first few accounts of ncRNA came from the *Bithorax* Complex (BX-C) in *Drosophila* (Grimaud, Negre, and Cavalli 2006, Lipshitz, Peattie, and Hogness 1987, Cumberledge, Zaratzian, and Sakonju 1990). *Drosophila* embryogenesis is characterized by a segment restricted expression of three protein coding genes from the *BX-C* each guided by cis regulatory elements present flanking the coding genes that are themselves transcribed in a time and tissue specific manner (Maeda and Karch 2006). Recent research has been more informative about the existence and possible role of ncRNA in mammals. Genome Network and FANTOM3 (Functional Annotation of the Mouse) projects co-guided the identification of all possible transcription units in mouse and humans using high throughput sequencing of 5'- and/or 3'- ends of transcripts from various tissues based on three approaches- 1) CAGE or capped analysis of gene expression, 2) GS or gene signature cloning and 3) GIS or gene identification of signature (Hayashizaki and Carninci 2006, EngstrÃ¶m et al. 2006, Carninci et al. 2005). This approach has led to the identification of cis-antisense pair of transcripts, bidirectional promoters and chains of transcriptional units bringing forth an updated version of the mammalian transcriptome (Carninci et al. 2005). Indeed, examination and comparison of tissue specific expression of coding and non-coding genes by other groups has led to a similar observation regarding the orientation of the two classes of genes (Mercer et al. 2008, Dinger et al. 2008). Non-coding RNAs can be classified into various classes based primarily on their pathway of genesis, length and known function. The different ncRNA identified to date are either short RNA (19 nucleotides to 30 nucleotides in length) or

long RNA (80 nucleotides to 40kb). Various classes identified to date are the yeast cryptic unstable transcripts and stable unannotated transcripts, mammalian promoter upstream transcripts, promoter associated long and short RNA, termini associated short RNA, miRNA, siRNA, piRNA and long non-coding intergenic RNA (ncRNA or lincRNA).

As mentioned earlier, RNA molecules of more than 80 nucleotides in length are classified as long intergenic ncRNA (also called lincRNA). The cut-off set for a valid open reading frame by the RIKEN group for ncRNA is < 300 nt in length which coincides with the observation that 95% of the proteins deposited in the Swiss-Prot and International Protein Index are more than 100 amino acids in length. The other criteria that distinguish ncRNA from protein coding messages are the degree of conservation, homology between species for the location within a given locus, conservation of the transcription start and end sites, splice site location and intron-exon architecture (Dinger et al. 2008a). With the advent of high throughput and global studies performed in order to detect the occurrence of transcription, isolate and analyze protein binding and tissue specificity together has led to the identification of a role for ncRNAs in various cellular processes including the regulation of homeotic gene expression, function as oncogenes, regulation of metabolic genes, regulation of skeletal development, eye development, epithelial to mesenchymal transition, organization of sub cellular structures etc. (Mattick 2009). Of relevance to the collinear and strictly regulated expressions of *Hox* genes during embryogenesis is the possibility of an involvement of ncRNA in defining the segmental domains within an embryo. *Abdominal B* in *Drosophila* represents the

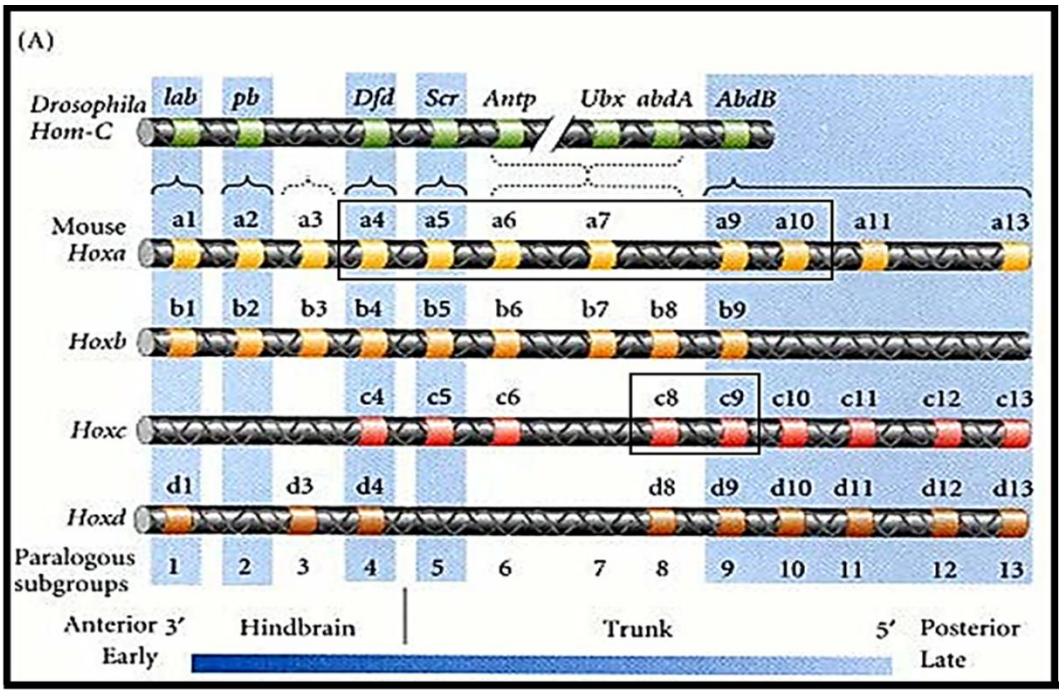
posterior most expressed gene that is regulated by iabs (infra abdominal regions) 5 through 8 in the corresponding parasegments (Bae et al. 2002). Testing for the presence of transcription reveal ncRNAs arising from the above mentioned iabs whose expressions occur in the same embryonic domains as that of *Abd-B,* however, the expression is observed to occur in a preceding manner possibly marking the domains for gene expression (Bae et al. 2002). As mentioned earlier, CMMs are comprised of both TrxG and PcG binding sites and are receptive to both activation and repression. *Fab7* represents one such CMM and a transient activation of it using a transgenic promoter early during embryogenesis leads to ectopic expression of the reporter gene in tissues with an otherwise silent *Fab7*. This effect is not observed when the transcriptional pulse is provided later during the larval stage denoting that the precise timing of activation through an otherwise silent element is critical (Rank, Prestel, and Paro 2002). Mammalian TRE/PREs have not been fully characterized yet and hence studies are lacking that could address the role of these elements and the maintenance proteins in *Hox* gene regulation. Recently, one study investigated *Hox* gene expression arising from all the four clusters in human tissues harvested from different regions along the anterior-posterior axis (Rinn et al. 2007). This microarray-based capture of gene expressions also led to an identification of many non-coding intergenic transcripts previously not reported in humans. Of particular significance to the identification of a role for ncRNA was the discovery of HOTAIR (HOX Antisense Intergenic RNA) transcribed from the *HOXC11-HOXC12* intergenic region that was found to repress the entire *HOXD* locus in trans via its interaction with the PRC2 group of repressive protein Suz12. This is a first account in

humans of a direct involvement of non-coding intergenic transcripts in the suppression of *HOX* genes in trans. Further characterization of HOTAIR by the same group has led to the finding that it binds to both the PRC2 complex as well as to the LSD1/CoREST/REST complex and performs a repressive function (Tsai et al. 2010). Another example of a regulatory long intergenic non-coding RNA came from the same group that identified a ncRNA HOTTIP from the 5' tip of the *HOXA* cluster, which supports expression from the 5' *HOXA* genes by binding to the WDR5/MLL complexes, thereby playing an important role in the recruitment of the co-activator protein complexes to the *HOXA* complex (Wang et al. 2011). These elegant studies represent recent attempts made at elucidating a role for ncRNA in mammalian gene regulation. Regulation of *HOXC8* by MLL has been demonstrated using both genetic as well as molecular means (Hanson et al. 1999, Yu et al. 1998, Milne et al. 2002). The MLL protein is capable of binding to co-repressor proteins (Xia et al. 2003) and this interaction is enhanced by the binding of CYP33 to the MLLPHD3 causing gene repression of MLL target genes (Fair et al. 2001). CYP33 thus functions as a polycomb group protein contributing repressive mechanisms at the MLL target genes. CYP33 indeed enhances the Polycomb mutant phenotype in *Drosophila* (Andrew Dingwall and Manuel O. diaz, unpublished data). It can thus be surmised that in early lineage cells, *HOXC8* may be bound by MLL in the presence of co-repressors such as the PRC1 group of proteins in a non-expressed state. In tissues where *HOXC8* is programmed for expression, mechanisms must exist that either inhibit the functions of co-repressor proteins or cause the removal of repressor proteins from the target gene.

Drosophila and mammalian _HOX_ gene clusters including the MLL target _HOXC8_, its regulatory elements and the implications of long range enhancer-promoter interactions.

Mammals and *Drosophila* contain homeobox genes that show a great degree of homology in sequence as well as their arrangement in the clusters in which they are found [Fig. 4]. Two *Hox* clusters have been identified in *Drosophila* namely the *Antennapedia* Complex (ANT-P) containing five protein coding genes and the *Bithorax* Complex (BX-C) that has three protein coding genes and both clusters are present on the chromosome 3R (right arm) (Duncan 1987, Lewis, 1954, Lewis 1963, Kaufman, Seeger, and Olsen 1990, Scott 1987) . Mutations that mapped to non-coding regions cause homeotic transformations similar to the phenotypes resulting from mutations in coding genes and were originally identified as pseudoallelic mutations (LEWIS 1951). These are now broadly recognized as mutations in the enhancer elements, which promote gene expression in a domain specific manner or in the boundary elements, which prevent enhancer-promoter interactions in the wrong expression domains (Gyurkovics et al. 1990, Mihaly et al. 1998, Hagstrom, Muller, and Schedl 1996, Zhou et al. 1996).

Figure 4: Shown in this adapted diagram are the *Drosophila* Antennapedia and *Bithorax* complexes in comparison with the mouse *Homeobox* clusters (A through D). The paralogous subgroups between *Drosophila* and mouse are highlighted. *Hox* genes follow the colinearity rule in that the 3' genes are expressed earlier than the more 5' genes with the expression beginning at the posterior end of an embryo and spreading more anterior during the course of development.

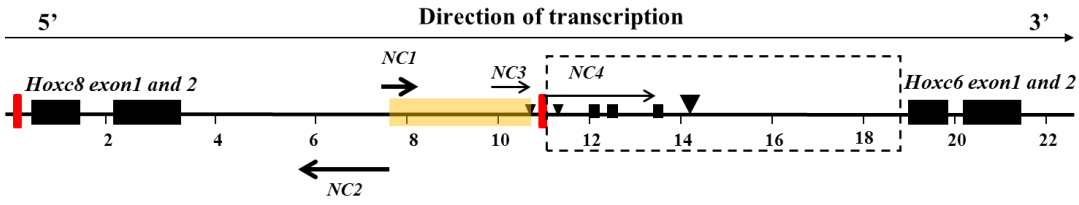


Developmental Biology, Scott F. Gilbert, sixth edition

There are 39 clustered *HOX* genes in human and mouse located in four different clusters namely the *HOX A, B, C* and *D* [Fig. 4] (Duboule 1998). Enhancer or inhibitory elements for certain genes have been identified e.g. regulation of the *HOXD* genes is under the control of both local cis regulatory elements as well as two global regulatory elements present flanking the entire cluster (Sabarinadh et al. 2004, Tschopp et al. 2009). It is believed that the outlying global regulatory elements control the expression of *HOX* genes early during embryogenesis when the segmental plan is being laid down within a developing embryo. The local regulatory elements are present within the cluster and influence gene expression in a tissue and cell type specific manner throughout adulthood (Tschopp et al. 2009). *HOXC8* present on human chromosome 12 and mouse chromosome 15 has been identified as a direct target of MLL. The expression of *HOXC8* has been well studied during mouse embryogenesis (Kwon et al. 2005) and is known to be biphasic; comprised of an early initiation phase and a late maintenance phase (Kwon et al. 2005). Genetic studies revealed a 5' regulatory element responsible for the initiation phase of *Hoxc8* during embryogenesis whereas a broad 8 kb region [Fig. 5] at its 3' end and 11kb downstream of the *Hoxc8* TSS is required for the maintenance of its expression in specific embryonic domains (Bradshaw et al. 1996). In addition to the requirements of the cis-regulatory elements, maintenance of *HOXC8* gene expression calls for a role of MLL protein too (Yu et al. 1998, Hanson et al. 1999). Expression of *HOXC8* occurs in all hematopoietic cells including the lymphoid, myeloid, erythroid and megakaryocyte cell types (Bijl et al. 1998). Similar to many clustered genes, the *HOXC* locus is comprised of cis-regulatory elements in the intergenic regions some of which, give rise to transcripts

identified as ESTs [Fig. 5]. In addition to the presence of transcripts, the *Hoxc8-Hoxc6* intergenic region contains various other features identified in different studies that support the hypothesis of its active involvement in gene regulation [Fig. 5]. Four ESTs were mapped in the region between *HOXC8* and *HOXC6* of which, *NC1* shows a greater degree of homology between diverse species ranging from *Fugu rubripes* through *Homo sapiens*. *NC1* was identified in human embryos as a transcript present about 7.6 kb downstream of the *HOXC8* transcription start site (TSS) and that splices into the exon 1 of each of the next three downstream genes such as the *HOXC6*, *HOXC5* and *HOXC4* (Boncinelli et al. 1989). As mentioned earlier, an 8 kb region about 11 kb downstream of the *HOXC8* TSS was identified in an initial genetic screen for regulatory elements controlling the maintenance of *HOXC8* expression during mouse embryogenesis (Bradshaw et al. 1996). Furthermore, three 400 bp long enhancers were identified in the region between 12 kb and 14 kb relative to the *Hoxc8* TSS (Milne et al. 2002). These enhancers were identified using a transgenic reporter assay performed in mouse embryonic fibroblasts that used LacZ tagged *Hoxc8* exon 1 with the promoter present downstream of the elements identified in the *Hoxc8-Hoxc6* intergenic region. Incidentally, all the three enhancers overlap the previously identified 8 kb *Hoxc8* 3'RR (Milne et al. 2002). In addition to the enhancers for *Hoxc8*, this study identified two minor and a major DNaseI hypersensitive sites in the intergenic region of which the major DNaseI hypersensitive site falls in the 8 kb long intergenic region (Milne et al. 2002).

Figure 5: *Hoxc8-Hoxc6* intergenic region is homologous between mouse and human and consists of features that are suggestive of its possible regulatory role. Shown in the schematic below is the *Hoxc8* gene with its 3' neighbor *HOXC6*. The *3' Hoxc8* RR is denoted by a rectangle about 11 kb downstream of the *Hoxc8* transcription start site (TSS). Three 400 bp enhancers for *Hoxc8* are shown by black boxes whereas the triangles represent the DNaseI hypersensitive sites identified (big and small triangles depict the major and minor sites respectively). The yellow shadowed region represents the bivalent chromatin identified for this region in mouse embryonic stem cells (Bernstein et al. 2006, 315-26). The red rectangles show the presence of clusters of 3 to 4 predicted YY1 binding sites. Importantly, the intergenic region is characterized by the presence of ESTs denoted here as *NC1*, *NC2*, *NC3* and *NC4*. The occurrence of these transcripts has been confirmed both in mouse and human cells in this study. *NC1* and *NC2* transcripts partially overlap. *NC3* is another transcript present just outside of the identified 3' regulatory region of *Hoxc8*. *NC4* is a 2.8 kb long transcript that has not been reported to undergo any splicing and partially overlaps with the *Hoxc8* 3' RR.



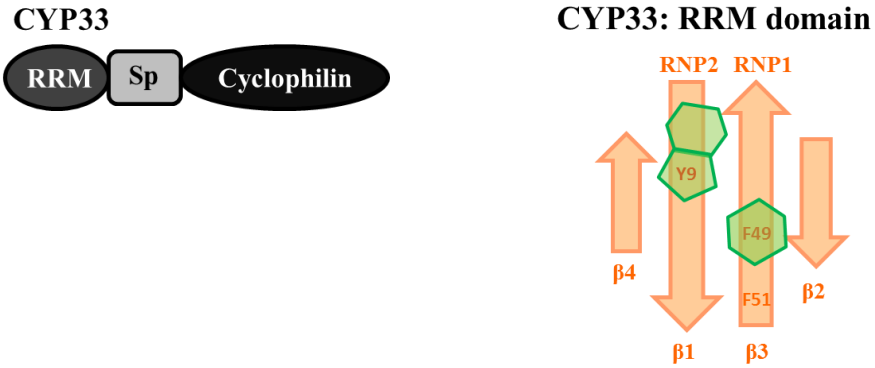
These results indicate that this intergenic region generates non-coding intergenic transcripts (ncRNAs). The 3' regulatory region of *Hoxc8,* may participate in enhancer-promoter interactions with the regulatory elements located in the intergenic region, to regulate tissue specific gene expression. Long range enhancer-promoter interactions have been a proposed mechanism for explaining the distance and orientation independent role of enhancer elements in the regulation of their associated promoters. Evidence in support of this hypothesis have been accumulating only recently for both Drosophila and mammals (Kagey et al. 2010, Nolis et al. 2009, Ronshaugen and Levine 2004, Tolhuis et al. 2011, Cai, Arnosti, and Levine 1996, Duboule and Deschamps 2004, Barna et al. 2002). Whether or not such long range interactions between *HOXC8* promoter and intergenic enhancers are mediated via the MLL complex remains an open question. Transcription of the 8 kb long *3'* RR into a ncRNA named *NC4* [shown in Fig. 5] may have a role in the regulation of *Hoxc8*. Presence of a bivalent chromatin structure in the 7.6 kb- 10.6 kb region [Fig.5] which is just upstream of *Hoxc8 3'* RR as well as *NC4* is suggestive of a regulatory role of this putative enhancer region in early lineage cells.

Modulation of MLL function towards gene repression via its interaction with CYP33

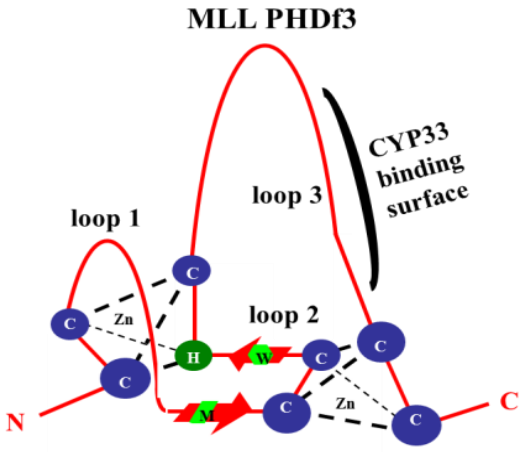
The third MLLPHD3 [Fig. 6a] from the MLL PHD finger cassette was initially found to interact specifically with a cyclophilin called CYP33 in a yeast two hybrid assay followed by confirmation of their direct interaction in a GST pull down assay (Fair et al. 2001, Anderson et al. 2002). CYP33 also co-localizes with MLL and exhibits a speckled distribution within the nucleus (Fair et al. 2001). CYP33 is a 33 kDa protein comprised of a 85 amino acids long N terminal RRM (RNA Recognition Motif), a 21 amino acids long conserved spacer and 195 amino acids long C-terminal cyclophilin domain [Fig. 6b]. The cyclophilin domain has a Cyclosporin A-sensitive Peptidyl Prolyl cis- trans isomerase (PPIase) activity (Mi et al., 1996).

Figure 6: CYP33 interacts with the MLLPHD3 via its RRM. (a) CYP33 is a nuclear cyclophilin with an N terminal RRM, a conserved spacer and a C terminal Cyclophilin domain. Shown in the adjoining panel is the RRM of CYP33 comprised of four antiparallel β sheets and containing the conserved RNP2 (Ribonucleoprotein 2) and RNP1. The second residue of RNP2 (Y9) and the fifth residue of RNP1 (F49) make stacking interactions with nucleic acid bases (shown are hypothetical purine and pyrimidine residues in green overlapping with the CYP33 β1 and β3 sheets). (b) The MLLPHD3 is a C4HC3 type zinc coordinating domain that binds CYP33 RRM via its loop3 as shown.

(a)

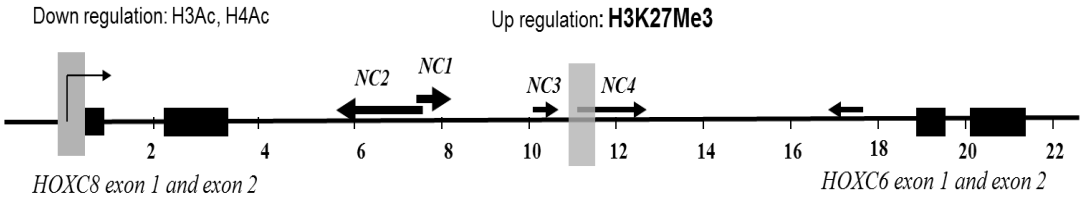


(b)



Its interaction with the MLLPHD3 occurs via the RRM and two independent studies have identified amino acids from this domain that are involved in this interaction (Park et al. 2010, Hom et al. 2010). Both groups identified amino acid residues I35, L39, E42, R47, F49 and F54 from the RRM involved in an interaction with the MLLPHD3 based on the NMR shifts observed in the backbone amides. On the other hand, amino acid residues L1571, G1589, E1605, M1606, Y1607, N1612, L1613, V1617, T1620 and E1626 from loop 3 of the MLLPHD3 are involved in making contacts with the RRM (Park et al. 2010). Furthermore, the two proteins interact with a dissociation constant in the range of 2-15 μM as determined by the two groups above (Park et al. 2010, Hom et al. 2010). The MLLPHD3 has been shown to interact with H3K4me3 also in vitro with a Kd of 30 μM (Park et al. 2010). This interaction; however is exclusive of its binding with the RRM (Park et al. 2010). Overexpression of CYP33 in the K562 leukemia cell line causes down regulation of *HOXC8*, which is also found to be dependent on the presence of MLL as well as both the RRM and the cyclophilin domains of CYP33 (Fair et al. 2001). The gene repression of *HOXC8* after CYP33 overexpression may occur at the level of chromatin control of transcription since overexpression of CYP33 in the 293 HEK cell line potentiates the recruitment of histone deacetylase 1 to the repression domain of MLL (Xia et al. 2003). Supporting this is the observation that upon CYP33 overexpression, acetylation levels at the histones H3 and H4 decrease thereby causing a down regulation of the *HOXC8* expression (Mark Koonce, Ph.D Dissertation).

Figure 7: CYP33 interacts with MLL to convert it into a transcriptional repressor of *HOXC8*. Shown below is a diagram of the *HOXC8-HOXC6* intergenic region showing the regions where histone deacetylation and H3K27 methylation have been observed after over-expression of CYP33 in a 293 cell line inducible for FLAG-CYP33. H3 and H4 deacetylation at the *HOXC8* promoter shown by shaded grey rectangle (Mark Koonce; unpublished data). An increase in H3K27me3 levels is also seen at the putative promoter of *NC4* also shown by shaded grey rectangle (Steven Poppen; unpublished data). The intergenic RNAs are indicated by arrows whereas the exons of protein coding genes are denoted in black rectangles.



Genetic evidence is available in *Drosophila* supporting the observation that CYP33 functions as a Polycomb group protein (Dingwall and Diaz, unpublished data). Addition of the MLLPHD3 alone or the three PHD cluster into an MLL-ENL fusion protein, which naturally lacks this domain, leads to an attenuation of transformation by the fusion protein (Chen et al. 2008). A separate study demonstrates that inclusion of the MLLPHD2, MLLPHD3 or the entire PHD cassette can attenuate the transformation of hematopoietic progenitor cells by MLL-AF9 (Muntean et al. 2008). These studies suggest that CYP33 may be an important link between the MLL complex of transcriptional activators and the co-repressor complexes including the PRC1 components (Steven Poppen; unpublished data) and that this interaction with the repressors may be important in order to maintain the proper levels of transcription through the cell divisions.

CYP33 also binds RNA via its RRM

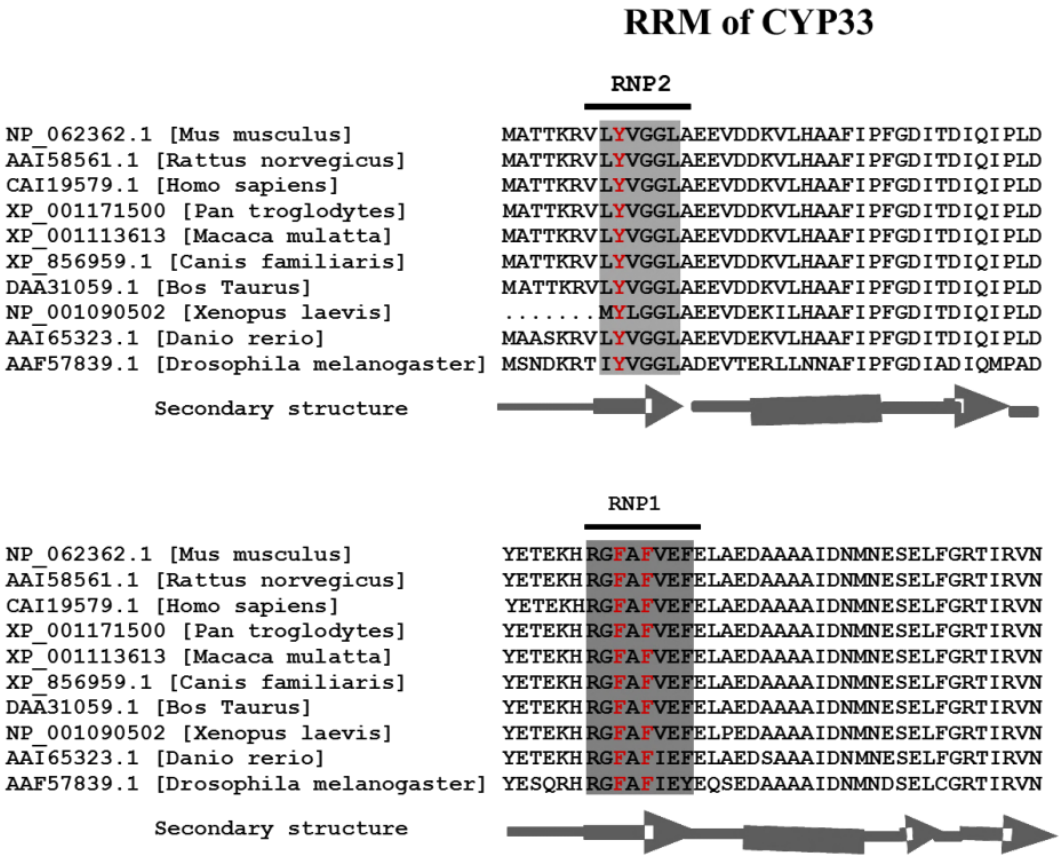
The RNA Recognition Motif (RRM) is an ancient conserved region that diversified by gene duplication. RRMs are in general capable of independently binding to the RNA, however, in some cases synergy between RRM and other protein domains is required for either general or sequence specific binding (Birney, Kumar, and Krainer 1993). The RRM can bind proteins and DNA in addition to RNA (Maris, Dominguez, and Allain 2005). Depending on the individual protein, binding between multiple partners may be simultaneous or exclusive e.g. RRMs from proteins like U2B and CBP20 bind both RNA and protein at the same time but need a cofactor for binding to the RNA (Maris, Dominguez, and Allain 2005). The CYP33 RRM was partially characterized by a group

that found it expressed in the Jurkat cell line (Mi et al. 1996). The CYP33 RRM preferentially binds single stranded Poly A and Poly U polyribonucleotides and with a very low affinity to ssDNA. It has a similar secondary and tertiary structure as the RRMs found in many other RNA binding proteins [Fig.1b] and forms a canonical β4α1β1β3α2β2 secondary structure with RNP2 (for Ribo Nucleo protein 2) placed on the β1 strand and RNP1 on the β3 strands. The four anti-parallel β strands are stacked against two α helices. Compared to the RRMs from other proteins, the β2β3 loop in CYP33 is extended and participates in specific molecular interactions with both RNA and the MLLPHD3. As is deduced for the canonical RRMs based on about 20 crystal structures available for comparison (Maris, Dominguez, and Allain 2005), the RNP2 and RNP1 each carry conserved aromatic amino acids that make stacking contacts with the bases from RNA [Fig. 6a] to form the primary RNA-binding residues of the RRM (Figure 6a shows the putative purine and pyrimidine bases in contact with the conserved residues of the CYP33 RRM). The RRM represents one of the most abundant and ancient protein folds found in organisms ranging from viruses through humans. Furthermore, an estimated 2% of all the proteins made in humans contain at least one RRM (Maris, Dominguez, and Allain 2005). Conventionally, the RNP2 and 1 fold to form a β sheet surface for interaction with RNA and only three conserved amino acid residues from the RRM participate in the interaction reducing the recognized core consensus to a dinucleotide (Maris, Dominguez, and Allain 2005). A dinucleotide is a very small consensus to confer any specificity to a CYP33 binding sequence. Amino acid residues outside of the RNPs, are also involved in RNA recognition e.g. residues from β2 and β4

participate in the identification of RNA thereby increasing the consensus to 4 nucleotides, which can be denoted as N_0-N_1-N_2-N_3 (Auweter, Oberstrass, and Allain 2006).The N_1 stacks with amino acid residue 2 of RNP2. The N_2 stacks with residue 5 of RNP1 [Fig. 6a]. The aromatic residue 3 of RNP1 inserts itself between the dinucleotide bases. In addition to these interactions, the N_0 is capable of interacting with either β4 or loop1-loop3 pocket. The N_3 however has a different mode of binding; it can either stack with N_2 to interact with the C terminal region of the RRM or it can independently stack with β2. Therefore a typical RRM has a tetranucleotide binding sites. In cases of RRMs where there are additional 3 nucleotides involved in binding, the loops 1, 3 and 4 participate in binding to these bases. The various intermolecular interactions responsible for ssRNA binding are pi-pi stacking interactions, cation- pi interactions and electrostatic interactions. The examples available so far suggest that nucleic acid recognition is a two-step process in which any RNA is attracted to the protein equally, however, if the stacking and H bond interactions that stabilize the interaction are not properly established, then the complex dissociates rapidly resulting in a weak affinity for the RNA oligonucleotides of non-specific sequence. Higher affinity is usually achieved by the presence of more than one RRM (Maris, Dominguez, and Allain 2005). CYP33 contains only one RRM and therefore may bind RNA with a moderate affinity (Mi et al. 1996, Hom et al. 2010). Shown in Fig. 8 are multiple alignments between CYP33 primary sequences from metazoans with RNP 2 and 1 containing the conserved RNA interacting residues. RRM of CYP33 binds both Poly A and Poly U polyribonucleotides (Mi et al. 1996) and the MLLPHD3 (Park et al. 2010; Anderson et al. 2002; Fair et al. 2001; Hom

et al. 2010). The third PHD (Plant Homeo Domain) finger of MLL is also a conserved zinc coordinating structure [Fig. 6b] that has been shown to be necessary and sufficient for binding to CYP33 via the RRM of the latter (Anderson et al. 2002; Fair et al. 2001). MLLPHD3 is a conserved CCCC-CCHC type zinc coordinating structure with a conserved tryptophan in the β sheet connecting the two zinc coordinating motifs, which is believed to be required for the interaction of MLLPHD 3 with H3K4me3 (Park et. al. 2010, Wang et. al. 2010). The interaction of CYP33 or trimethylated H3K4 with CYP33 is mutually exclusive of each other (Park et. al. 2010).

Figure 8: Multiple alignments of the primary sequence of CYP33 RRM from various species of metazoans. The RRM represents an abundant and conserved domain present in all life kingdoms ranging from viruses through humans. The RRM typically consists of RNP2 (highlighted grey) with the consensus I/V/L-F/Y-I/V/L-X-N-L and RNP1 (highlighted dark grey) with the consensus of K/R-G-F/Y-G/A-F/Y-V/I/L-X-F/Y where X is any amino acid. The invariant aromatic amino acids in RNP 2 and 1 are denoted in red. Secondary structure elements formed by the RRM are also shown.



Apart from binding to Poly A and U homopolymers, CYP33 was proposed to bind preferentially to the AAUAAA polyadenylation signal of a cellular mRNA pool (Wang et al. 2008). This study, however, has not identified a physiological significance of such an association. With the two known RNA binding sequences, there have been recent attempts at identifying the amino acid residues from the CYP33 RRM that bind Poly A (Park et al. 2010) or the AAUAAA motif (Hom et al. 2010). Following the convention for RRM, both these studies identified critical aromatic residues Y9 (and V10) from RNP2 and F49 and F51 from RNP1 that show displacement upon binding to either of the RNA sequences in an NMR assay. When compared, the binding surfaces on the RRM for RNA and MLLPHD3 overlap to a certain extent with the F49 residue involved in an interaction with both the molecules (Park et al. 2010, Hom et al. 2010). Since neither of the approaches above represents an unbiased method of RNA ligand identification for CYP33, the current study set out to identify a specific RNA sequence binding to CYP33 from amongst a pool of random RNA sequences in an in vitro selection assay of SELEX (Selective Evolution of Ligands by Exponential Enrichment) (Sakashita and Sakamoto 1994).

CHAPTER THREE

Materials and Methods:

In silico analysis of *HOXC8-HOXC6* intergenic region: Inter species multiple alignments were done using the NCBI's nBLAST and search for ESTs using the EST database for non-redundant RNA sequences from *Homo sapiens* and *Mus musculus* (GI numbers for known ESTs provided in Table 3). The entire region was compared between mammalian species for conserved features such as the promoter proximal and distal elements (Hoon-Le et. al. 2005, Yang et. al. 2007, Suzuki et.al. 2001), CpG boxes (Suzuki et. al. 2001), DNA transcription factor binding sites using the TransFac prediction algorithm (Matys et al. 2003). MLL binding sites and the bivalent marks on chromatin were identified by examining the data from literature on chromatin immunoprecipitation experiments (ChIP on chip array) (Bernstein et al. 2006) in global studies performed in mouse embryonic stem cells.

Prediction of protein coding potential for the transcripts found in *HOXC8-HOXC6* intergenic region: Protein coding potential of the *HOXC8-HOXC6* intergenic transcripts was estimated using Coding Potential Calculator (CPC) (Kong et al. 2007), NCBI ORF finder, conservation of sequence, conservation of position within the locus, presence of

promoter proximal elements, conservation of exon-intron architecture, a valid open reading frame (more than 300 nts.), potential splicing sites and polyadenylation signals. For protein translation potential, a valid mammalian Kozak sequence involved search for the consensus GCCRCCaugG ('R' is a purine and 'aug' is the initiation codon) (Xia et. al. 2007).

Cell culturing:

Cell lines:

The HEK 293 cell line was obtained from American Tissue Culture Collection (ATCC), maintained in Dulbecco's modified Eagle's medium (DMEM) with high fructose (Gibco BRL) and supplemented with a final concentration of 10% Fetal Bovine Serum (Gibco BRL) and 1% Penicillin/Streptomycin (Gibco BRL) at 70%- 80% cell density. Transfections, were performed using 60% confluent HEK 293 cells in a 6 well plate (Corning) or 10cm culture dishes (B D Falcon) and 4μg or 10μg DNA respectively (as per manufacturer's protocol) combined with the cationic lipid based transfection formulations (Invitrogen). Cells were harvested 24 hours or 48 hours after transfection without intervening medium change, for further analysis.

MEFs (both MLL wild type and MLL null generated by (Yu et al. 1995) were maintained at 80% - 90% cell density in 10 cm culture dishes (B D Falcon) with DMEM high glucose containing 10% FBS, 1% Penicillin/ Streptomycin and 0.1mM β-mercaptoethanol. Cells were harvested for total RNA using Trizol (Invitrogen) followed by the manufacturer's protocol for RNA extraction.

MSA cell line: The human thyroid carcinoma cell line MSA was kindly provided by (Dr. Shoji Nakamori, Osaka National Hospital, Japan) and cultured in DMEM F/12 (50%-50%) high glucose (GIBCO BRL), supplemented with 10% FBS and 1% Penicillin/Streptomycin at 80%- 90% cell density for routine maintenance.

Mouse embryos: Wild type mouse embryos from the three post gastrulation stages were obtained from Dr. Cooduvalli Shashikant (The Pennsylvania State University, PA) upon request. The embryos were provided in RNAlater® (Ambion) for the purpose of RNA extraction in RNA based studies.

In vitro differentiation of RW4 mouse embryonic stem cells into embryoid bodies: RW4 mES cells were kindly provided by Dr. Gerard Grosveld and cultured according to the protocol from Stem Cell Technologies (all the reagents were purchased from Stem Cell Technologies). Low density (3- 4 X 10^5 cells/ml) cells were frozen in 90% FBS (ES tested by Stem Cell Technologies) with 10% filter sterilized DMSO (Dimethyl Sulfoxide). Aliquots for each experiment were thawed overnight and grown in the DMEM high glucose medium with 15% FBS (ES tested by Stem Cell Technologies), 1% Penicillin/Streptomycin, 1 mM Sodium Pyruvate, 2 mM Glutamine, 0.1 mM Non-essential amino acids, 10 μg/ml of Leukemia inhibitory factor (LIF) and 1:100 diluted Mono Thio Glycerol (MTG from Sigma). Pre-differentiation was performed for no more than two days in Iscove's modified Eagle's medium (IMDM) containing all of the above reagents. Differentiation is induced by removal of LIF from the medium and plating cells at very low cell density (10,000cells/ml) in differentiation medium without LIF. The

culture dishes used for undifferentiated mES cells are coated with 0.1% filter sterilized gelatin. Differentiation induction medium is mixed as follows:
MethoCult M3234 (Stem Cell Technologies; thawed at 4°C overnight or more): 52.5 ml ES cult FBS (either EScult 6900 or 6902): 23.7 ml, IMDM: 134.3 ml, L-glutamine: 1.58 ml, MTG (1:100): 196 µl, filter sterilize and mix with differentiation medium (Stem Cell Technologies Cat # M3234). 60mm petri dishes are used for this experiment. This ensures that the cells do not settle on the bottom. Total number of cells needed per experiment is 2.1 X 10^6. RNA is extracted from EBs harvested every day using DMEM and centrifugation at 800 rpm/5 min.

RNA isolation and cDNA synthesis:

RNA isolation was performed according to the manufacturer's protocol using Trizol reagent. After the two-phase separation of RNA in an aqueous layer, it is precipitated using isopropanol:water (2:1) and washed twice with 75% ethanol. The pellet is dried for no more than 5 minutes and dissolved in an appropriate amount of Diethyl Pyro Carbonate (DEPC) treated water. The RNA solution is heated at 65° C and diluted 1:100 for spectrophotometric analysis at A_{260} for an estimation of the total amount. The required amount of total RNA thus purified is used for first strand cDNA synthesis (or other analysis).

First strand cDNA was synthesized (for most experiments) using 2 µg of total RNA and random hexamers from High Capacity cDNA Synthesis kit (Applied Biosystems). The

synthesis is done at 37°C for 2 hours following which the 20 µl of cDNA reaction mixture is diluted according to the requirements and used for PCR analysis.

Semi quantitative and quantitative RT PCR:

20 µl of cDNA is diluted 1:2 using 10 mM Tris (pH 8) and stored at -20°C as a main stock. This main stock cDNA is further diluted depending on the mRNA to be detected. All the semi-quantitative PCR were performed using 1 µl or 2 µl of the main stock in GoTaq green master mix (Promega) and primers of interest. Quantitative PCRs were performed using Promega's GoTaq qPCR master mix and either 1:6 diluted cDNA for high abundance RNA, 1:10 for housekeeping RNA and undiluted for low abundance transcripts. The primer amount was set to a final concentration of 0.5 µM for semi-quantitative PCR and 1.25 µM for quantitative PCR analysis of low abundance transcripts and 2.25 µM for high abundance transcripts including the housekeeping transcripts. Annealing temperature for semi-quantitative analysis ranged from 55°C-60°C depending on the primer pair (see table 2 for primer sequence). For quantitative PCR, the annealing temperature was set to 60°C.

miRNA detection:

1) **Denaturing gel for nucleic acids-** 7M Urea (denaturing) polyacrylamide gels polymerized using a mini gel apparatus. 15 ml of 7M urea gels are prepared to run on the X cell sure lock mini cell (BioRad) as follows: 6.3g urea, 5.625 ml, 40% Bis-Acrylamide

solution, 3 ml 5X TBE and DEPC treated water to make up the total volume to 15 ml. The components were mixed and the solution warmed to 37°C in order to dissolve the urea. The gel was then filtered through a nitrocellulose filter and cooled to room temperature. 90 µl of 10% APS (fresh) and 7 µl of TEMED are added and the gel solution is mixed thoroughly before pouring into the mini gel apparatus avoiding air bubbles. 1 mm thick combs were used and after the gel was set, the comb was removed and the wells rinsed with 0.5X TBE. Place the plate into the gel apparatus (containing 0.5X TBE).

2) **Sample preparation and resolution on the gel**. 2X RNA-loading dye (8M Urea, 20 mM EDTA, 0.1% Xylene Cyanol and 0.1% bromophenol blue or BPB) prepared by mixing Urea 8.4g, EDTA (0.5M pH 8): 800 µl, Xylene Cyanol 20 mg, BPB 20 mg, DEPC treated water to a final volume 20 ml. The RNA is dissolved in the minimal possible volume of DEPC treated water and its concentration estimated from the A_{260}. RNA is diluted with 2X loading dye in order to obtain a 1X RNA sample. After dissolving RNA in 1X (8M Urea) loading dye; the samples are heated at 80°C for 5-10 min. Samples are then spun down and loaded onto the urea free wells. The gel is then run at 200V for 1hr. The voltage is then raised to 500V for 1.5 hr or until the BPB just runs off the gel. The gel is placed onto Saran wrap and stained with 4 µg/ml EtBr in 0.5X TBE/ 5 min. RNA is transferred onto Nitrocellulose membrane using electro-transfer in the BioRad Mini Protean system transfer apparatus using 0.5X TBE at 100V for 1hr. Complete transfer is ensured by absence of ethidium bromide positive bands in the gel.

Wet Nitrocellulose membranes were cross linked using the Biorad, GS gene linker UV chamber at 150 mJoules for 30 sec.

3) **DNA or RNA probe hybridization.** 5' end-labeled DNA probes were used for hybridization (for both *NC1* DNA strands). The DNA primers were ordered from Invitrogen and were provided as non-phosphorylated single stranded DNA oligonucleotides (see table 2 for sequences of hNC1-F and R) that could be labeled at the 5' end using the T4 Polynucleotide Kinase (Fermentas). Standard protocol from the manufacturer (Fermentas) was followed for the labeling. The probe was dispensed directly into the Hybridization Solution after free nucleotide removal using G50 columns (GE Healthcare).

Hybridization temperatures: DNA: RNA hybrids:

Tm= 79.8 C + 18.5 (log10 [Na+]) + 0.58 (%G+C) +11.8(%G+C) 2-0.50(% formamide)-(820/1).

Before the hybridization step; the membrane is treated as follows:

The membrane carrying the immobilized RNA is hydrated in 6X SSC. It is then placed with RNA side up in a hybridization tube and 1 ml of formamide containing prehyb./hybridization solution per 10 cm^2 of membrane is added. The tube is then placed in the hybridization oven and incubated at a set temperature with rotation for 3 hours or more. The desired volume of the probe is pipetted into the hybridization tube and continued to incubate with rotation overnight at a preset temperature. The membrane is then washed as follows after the hybridization step-

2X SSC/ 0.1% SDS incubated with rotation for 5 min at room temperature, the wash solution changed and repeated. The next washing steps involve the following steps in series: First prepare the following solutions- Wash 1: 2X SSC, 0.5% SDS; Wash 2: 2X SSC, 0.1% SDS and Wash 3: 0.1X SSC, 0.5% SDS. The washes are performed as- Wash 1/ RT/ 5min, Wash 2/ RT/ 15 min, Wash 3/ 37° C/ 1hr, Wash 3/ 68° C/ 1 hr, Wash with 0.1X SSC. The membrane is placed in a sealable bag expose to X ray film. The exposure time should be determined empirically.

SELEX: Cloning of cDNA into appropriate bacterial expression vectors, protein expression and purification followed by the SELEX protocol as described below:

Cloning of various cDNAs for protein expression- In order to obtain hCyp33 without the RRM, PCR mediated cDNA amplification was performed using primers with EcoRI and NotI sites (hCypRTEcoF and hCypRTNotR; see table 2 for sequences) and cloned into pGEMT Easy followed by sub cloning into pGEX4T-1 downstream and in frame with the GST coding sequence. The full-length hCyp33 cDNA was excised from the EcoR I and Not I sites of a vav-hCD4-hCyp33 construct (made by Wei Wei) and cloned at the same sites into the pGEX-4T-3 vector (Amersham Biosciences) downstream and in frame with the GST coding sequence.

(I) Protein expression in bacterial system and purification:

Various GST protein constructs were transformed into BL-21 competent cells (Stratagene) for large-scale protein expression using standard protein purification

protocol for GST tagged proteins. For preparation of GST tagged proteins on a large scale, 800 ml Luria Broth (LB) containing 100μg/ml of Ampicilliin was inoculated using 80 ml of overnight grown culture of BL-21 cells transformed with given constructs. The inoculated 800 ml LB is allowed to incubate at 37°C with aeration until the optical density (A_{600})reaches 0.6. Following this, the protein expression is induced using 1mM IPTG (Isopropyl β-D-1- thiogalatopyranoside from Sigma) and the culture is allowed to express protein at room temperature with aeration for 4 hours or overnight. The bacteria are then pelleted using Beckman Centrifuge at 6000rpm for15 min at 4°C and either stored at -80°C or processed further for protein extraction. The bacterial pellet is suspended in Saline Tris EDTA (STE) buffer (10mM Tris at pH 8.0, 150mM NaCl) containing 100 μg/ml of lysozyme and incubated on ice for 15min. To this, a final concentration of 5mM DTT (Dithiothreitol from Sigma), 1X mini protease inhibitor cocktail tablet (Roche) and 1.5% final concentration of N-laurylsarcosine are added followed by vortexing for 5 seconds. The supernatant is adjusted to final concentration of 2% Triton X and mixed well. Adequate lysis is ensured by performing either sonication or using French Press (each lysate pressed thrice or more to ensure complete lysis). The lysate is centrifuged using the Beckman Centrifuge at 10,000rpm for 30 min at 4°C and the supernatant is incubated with 1 ml of packed volume of STE-washed Glutathione Sepharose 4B (GE Healthcare) over night at 4°C on a spin-wheel. The beads are then washed with washing buffer (20mM Tris at pH 8.0, 120mM NaCl, 10% glycerol, 0.1% Triton X and 1X mini protease inhibitor cocktail) for 5 times or more using 4 times the bead volume for each wash. Following washing, the bound protein is eluted from the

beads using excess reduced glutathione in elution buffer (10mM Tris at pH 8.0, 50mM reduced glutathione and 10% glycerol). Further purification was done using size exclusion filter columns from Millipore and the purified protein was stored as aliquots in 50 mM Tris, pH 8.0 containing 10% glycerol at - 80°C.

(II) Generation of a random RNA pool:

Double stranded DNA template containing a T7 polymerase promoter site followed by a 30 nucleotides long randomized region and a 3' constant primer binding site was generated by using the following primers (ordered as single stranded DNA oligonucleotides from Invitrogen):

T7Univprimer2:5'CATCTGCAAGTACTAGAGTAATACGACTCACTATA*G*GACTG ACCTAGTCTGAC 3', Biotin-RevUnivPrimer2: 5' Bio/CTGACACTGCAGTCTGAG 3', LinearN30: 5' CTGACACTGCAGTCTGAG**(N_{30})**GTCAGACTAGGTCAGTC 3', Biotin SlxHeel: 5' Bio/CATCTGCAAGTACTAGAG

Preparation of ds Transcription template (in standard 1X PCR buffer): 5ng of Linear N30, 5ng of T7Univprimer2, 1ug of Biotin- RevUnivPrimer2, 1ug of Biotin- SlxHeel, PCR at 94/4min, [94/1min, 54/1min, 72/1min (10X)], 72/2min.

In vitro transcription using T7 polymerase reaction (Fermentas): PCR products run on 1.2% agarose gel extracted and eluted in a 20 µl elution buffer (10mM Tris pH-8.0, 1mM EDTA). 3 µl used in an in-vitro transcription reaction using the Mega Shortscript kit (Ambion) and incubated for 5 hrs at 37°C to give a final volume of 20 µl. The reaction was diluted using 360 µl DEPC water and 20 µl of 50% Streptavidin agarose slurry (Novagen) was added to each reaction to remove the biotin labeled DNA PCR product

(4°C/30min). The reaction was then centrifuged at 4°C/ 14000rpm/ 10min and supernatant removed followed by precipitation using 10% v/v of 3M Na acetate and twice the volume of 100% Ethanol. The RNA is suspended in 10 µl DEPC water and heated at 65°C for 5min. Set up DNase I (Fermentas) digestion at 37°C for 15min. Terminate the reaction at heating at 65°C for 5min followed by one step Phenol: Chloroform: Isoamylalcohol extraction and finally by salt/ethanol precipitation. Preclearing using GST Glutathione beads was done as follows: 20 µl of RNA pool (post DNase I treatment; Streptavidin pull down), 2 µl or 80 units of RNasin, 10 µl of Protease inhibitor (10X), 25 µl of 4X SELEX buffer, 10 µl of Glutathione Sepharose 4B with bound GST, 33 µl of DEPC water to a total of 100 µl. Prebinding or clearing were done at 4°C for 1hr and the reaction was centrifuged to remove the beads.

Binding between GST-CYP33 and random RNA pool:

The supernatant is used as a random RNA pool for binding at 4°C for 1hour to the Glutathione Sepharose 4B bound GST-CYP33 in a 1X Binding buffer containing 200 mM HEPES-NaOH pH 7.9, 200mM KCl, 5% Glycerol, 1X protease inhibitor cocktail (Roche). The beads are then washed 5 times with 400 µl of binding buffer and eluted in 50 µl of 1X elution buffer containing 10mM Glutathione and 50mM Tris-HCl, pH 8.0 at room temperature/15 min. The protein was removed using Phenol:Chloroform:Isoamyl alcohol (25:24:1) followed by a salt/ethanol precipitation of the eluted RNA. 20% of the recovered RNA is reverse transcribed into the first strand cDNA synthesis using random hexamers as described in the first step. PCR is used to generate the dsDNA template which, following the in vitro transcription step, is used in the second round of binding to

GST-CYP33. After the final round of binding, the dsDNA thus generated is T/A cloned into pGEMT-Easy cloning vector (Promega) and sequenced using the following primers (pGEXupF: 5'AGTATATAGCATGGCCTTTGAAG 3', hCYP33midSeqR: 5' TTTGTGAAATCACCGCCCTG 3' and pGEXupR: 5' ACAGACAAGCTGTGACCGTCTCC 3'). The DNA sequences thus generated are aligned for motif search.

(III) Data analysis for SELEX: Multiple alignment algorithms currently available over the web are inadequate to detect small nucleic acid sequence consensus. Therefore, the selected nucleic acid sequences were corrected for 5' to 3' orientation and used for both primary sequence analysis and prediction of secondary structure. The primary sequence analysis involved detection of high frequency occurrence of all possible dinucleotide combinations and increasing the consensus in both directions from this seed sequence. The motif found was compared with the sequences selected by the control protein without the binding domain for RNA (GST-ΔRRM CYP33) as well as a point mutant that does not bind Poly A in vitro [Fig. 14a] (GST- CYP33L72P). Frequency of occurrence was calculated as follows:$[(P)^n]$ x L, where 'P' is the probability of occurrence of a base (0.25 for A/U/G/C, 0.5 for a R or a Y and 1 for N where 'R' is a purine base, 'Y' is a pyrimidine base and 'N' is any of the four nucleotides, 'n' is the number of positions occupied and 'L' is the total length of the polymer (in this case 44 of the 30 nucleotide sequences selected by GST- CYP33 were considered as a single polymer of 1320 nucleotides). Chi-square distribution was done to test if the observed frequency was significantly different from the expected frequency. Secondary structure analysis

involved in silico folding of the selected sequences at 37°C using the mFold algorithm devised by the Zuker group (Zuker 2003, 3406-15) and available over the web (http://mfold.rna.albany.edu/?q=mfold). A single sequence could fold into more than one possible structure. Each structure was compared for available free energy, occurrence of whole motif (YAAUNY and AAU) within the loop, outside of the loop or in double stranded regions of the folded RNA (see table 3).

RNA electrophoretic mobility shift assays:

In vitro binding between GST-CYP33 and 5' end labeled RNA probes followed by resolution of the bound fraction from the free probe in a non-denaturing polyacrylamide gel.

1) GST-CYP33 was prepared from pGEX4T-CYP33 transformed DH5α *E. coli* cells using a standard protocol for GST tagged proteins. Further purification was done by molecular exclusion chromatographic columns (Amicon by Millipore, Ultracel- 3K). The protein was resuspended in 10mM Tris (pH 8.0) containing 10% glycerol and frozen at -80°C.

2) The probes were prepared by *in vitro* transcription of double stranded DNA templates (see table 2) containing a 5' T7 RNA polymerase promoter using the T7 Polymerase kit from Fermentas. After protein removal using the standard Phenol:chloroform:isoamylalcohol step, RNA was precipitated, its concentration estimated from the A_{260}, and used for 5' labeling in a forward reaction by the T4

Polynucleotide kinase (T4 PNK) from Fermentas and γ- 32P ATP (Amersham) with specific activity of 6000 Curie/mMol. Subsequently, the labeled RNA was filtered through G25 molecular exclusion matrix columns (GE Healthcare) to remove free nucleotides. The purified and labeled RNA thus obtained was used with GST-CYP33 in an in vitro binding reaction containing 1X binding buffer (25 mM Tris pH 6.7, 200 mM NaCl and 2.5 mM $MgCl_2$) at 37°C for 1 hour before resolution on a non-denaturing polyacrylamide gel.

3) An 8% native polyacrylamide gel was prepared by mixing 8 ml of (40%) Bis acrylamide, 5X TBE, 10% APS and 0.001% (v/v) TEMED.

4) The samples are loaded with 1X RNA native sample buffer and the gel is run at 100V until the required separation is achieved (Xylene Cyanol and Bromophenol blue are used as tracking dyes). The gel is dried in a Bio-Rad Model 583 gel drier before exposure to X ray films (Kodak) for imaging. Quantification, wherever applicable, was performed using Phosphor screens and a Typhoon Phosphor Imager (GE Healthcare Lifesciences).

FRET:

1) **cDNA cloning and protein expression:** Cerulean-CBFα and Venus-CBFβ cDNAs cloned into the parental vector pET22b+ were obtained from Dr. Bushweller (University of Virginia). Full length CYP33 was cloned into the parental vector pET22b in frame with the Cerulean coding sequence by first removing the inserted cDNAs of CBFα and CBFβ using the restriction endonuclease sites BamHI and XhoI and then inserting the

PCR generated CYP33 fragment (forward primer 5' GGATCCGCCACCACCAAGC and reverse primer 5' CTCGAGGCCTCACACGTACTC) at the same sites. Likewise PCR generated CYP33ΔRRM (forward primer 5' GGATCCGGCTCTTCCAGGCCAG and reverse primer 5' (CTCGAGGCCTCACACGTACTC) was cloned into the same sites in Cerulean containing pET22b+. MLLPHD3 cDNA was obtained by the restriction digestion of pGEX4T-1-MLLPHD3 using BamH1 and Xho1 restriction endonucleases and cloned in the pET22b+ vector in frame with the Venus coding sequence. The bacterial plasmid constructs thus obtained expressed proteins each with an N terminal and a C terminal Hexahistidine tag as well as an N terminal fluorescent (Cerulean or Venus) tag. For protein expression, BL21 (DE3) Codon Plus RIPL *E. coli* competent cells were transformed with each of the above constructs. Protein expression, extraction and purification of the proteins was performed using the standard protocol. The His-tagged protein constructs are used to transform the BL-21 DE3 Codon Plus RIPL competent cells and plated on LB plates containing 100 μg/ml Ampicillin (Amp) and 34 μg/ml Chloramphenicol (Cm). A single colony is re-streaked on the Luria agar plate containing amp (100μg/ml), Cm (34μg/ml) plate and grown overnight. A single colony is inoculated in 80 ml LB containing 100μg/ml of Amp, 34 μg/ml of Cm and grown overnight. An 800 ml LB containing 100μg/ml amp, 34μg/ml Cm is inoculated with 80 ml of the overnight culture and allowed to grow at 37°C under aeration until the optical density (A_{600}) of the culture reaches 0.6. At this point, a final concentration of 1mM IPTG is added to the culture for protein induction and incubated at room temperature under aeration condition from 4 hours to overnight. The cells are then harvested by spinning in the Beckman

centrifuge for 15 minutes at 6000 rpm. The pellet is then suspended in lysis buffer (10 mM Tris at pH 8.0, 100 mM sodium phosphate or NaH_2PO_4, 50 μM $ZnSO_4$, 300 mM NaCl, 10 mM imidazole, 1X mini protease inhibitor cocktail, 5mM β-mercaptoethanol, 1% Trithon X-100 and 5 mM DTT). Lysozyme is then added to a final concentration of 1μg/ ml (20 μl/ml of the stock) and incubated on ice for 30 min. Triton X-100 is added to a final concentration of 0.1% and mixed well. The lysate is then subjected to French press for complete lysis followed by centrifugation at 10,000 rpm for 30 min at 4°C to pellet the cellular debris. The supernatant is used for protein pull down by addition of up to 4ml (depending on expected yield of protein) of packed volume of Ni-NTA- agarose beads (Qiagen). The protein bound beads are washed using wash buffer (100mM sodium phosphate or NaH_2PO_4, 10mM Tris at pH 7.0, 500 μM $ZnCl_2$, 300mM NaCl, 1X mini protease inhibitor cocktail, 60mM imidazole, 5mM β-mercaptoethanol and 10% glycerol) for 5 times or more using 4 times the bead volume for each wash. The protein bound to the beads is then eluted using the elution buffer (10mM Tris pH 7.0, 100mM sodium phosphate or NaH_2PO_4, 50μM $ZnCl_2$, 300mM NaCl, 250mM imidazole, 5mM β-mercaptoethanol and 10% glycerol). The purified protein is resuspended in 10mM Tris (pH 8) and 10% glycerol after protein estimation using the BCA reagent.

2) **Optimization of the protein- protein interactions for the competition assays:**

FRET was performed using a 1X binding buffer (25mM Tris pH 8, 200mM NaCl, 2.5mM MgCl2, 1 ug/μl BSA, 1mM DTT) optimized for both protein-protein and protein-RNA interactions in our lab. A titration experiment was done using Cerulean-CYP33 or Cerulean-CYP33ΔRRM with increasing amounts of Venus-MLLPHD3 which yielded an

estimated Kd of 10.97 μM. The combination of the two proteins was set to 4 μM of Cerulean tagged CYP33 and 35 μM of Venus tagged MLLPHD3 in the competition assays by taking into consideration the observation that FRET efficiency increases when the ratio of Donor/Acceptor decreases (Berney and Gaudenz, 2003).

3) **RNA used for FRET based competitive binding assays:** RNA of different sequence composition were synthesized using dsDNA based templates carrying a T7 RNA polymerase site (Fermentas) using the manufacturer's protocol. The RNA thus generated was purified from the reaction using phenol: chloroform: isoamyl alcohol (25:24:1) and salt precipitation (3M sodium acetate) followed by further purification using the G25 Sephadex matrix based columns in order to filter out the free nucleotides. The filtered RNA concentration was estimated from the A_{260} and it was used for the binding at the same molar concentration as that of MLLPHD3 (35 μM).

4) Competitive interactions: The competition assays were performed using all the three components added in the amounts mentioned above in a 1X binding buffer and incubated at 37°C for 1 hr following which a 5 μl aliquot of each steady state binding reaction was tested for FRET signal using the 3cube FRET method (Hou, Kelly, and Robia 2008).

5) **3 cube FRET:** 3 channel excitation and emission was performed using the filters (430 nm/470 nm with bandpass width of 24 nm) for the Cerulean channel, (500 nm/535 nm with bandpass width of 24 nm) for the Venus channel and (430 nm/535 nm with bandpass width of 24 nm) for the FRET channel readings. Individual donor and acceptor proteins (both in the presence and absence of PolyA or free nucleotides, 35 μM) were

used in each experiment to correct for the channel bleed through. The positive control for FRET generated by protein-protein interaction involved binding between Cerulean-CBFα and Venus-CBFβ. RNA was added to this interaction to test whether the FRET signal disruption was specific for CYP33-MLLPHD3 interaction.

6) **Fluorescence imaging microscope equipped with quantitative data acquisition:** Fluorescence measurements were performed in polystyrene 96 micro well mini trays from Nalgene Nunc International using an inverted microscope (Nikon Eclipse Ti) equipped with focus drift correction, motorized excitation/emission filter wheels (Sutter Instrument Co., Novato, CA) and a motorized stage (Prior, Rockland, MA). Acquisition was automated with custom software macros from Meta-Morph (Molecular Devices Corp., Downingtown, PA). Metal halide lamp (Prior Scientific; Lumen 200) illumination was introduced through an excitation filter wheel equipped with 430/24 nm (for CFP), a 500/20 nm (for YFP) bandpass filters and a stationary multiple band dichroic mirror (Semrock, Rochester NY). Emission was collected with a 40X 0.75 N.A. Plan Fluor objective and an emission filter wheel equipped with 470/24 nm (for CFP) and 535/30 nm (for YFP) filters. Fluorescence emission was quantified with a back-thinned CCD camera (iXon 887; Andor Technology, Belfast, Northern Ireland) cooled to -100°C with recirculating liquid coolant system (Koolance, Inc., Auburn, WA).

7) **Calculations:** Normalized (sensitized) FRET was calculated as follows: Corrected FRET= Ida-aIaa-dIdd, where, Idd: Fluorescence intensity of donor channel for donor only, Iaa: Fluorescence intensity of acceptor channel for acceptor only, Donor spectral bleedthrough defined by aIaa and Acceptor spectral bleed through defined by

dIdd, where ‘a’ and ‘d’ refer to the bleed-through correction factors. The correction factor ‘a’ is calculated for acceptor only by taking a ratio of fluorescence intensity from FRET channel and fluorescence intensity from the acceptor channel. The correction factor‘d’ is calculated for donor only by taking a ratio of fluorescence intensity from FRET channel and fluorescence intensity from the donor channel. Finally, the Normalized FRET is calculated as corrected FRET/ (Corrected FRET+ 3.2 X Idd), where 3.2 is the G correction factor based on CFP-YFP pair. G is the ratio of sensitized emission to the corresponding amount of donor recovery after acceptor photobleaching, which was determined to be 3.2 for this set up.

Induction of *HOXC8* in MSA cell line using Cyclosporin A:

MSA cells were treated at 60% confluency in a 10 cm dish with Cyclosporin A (Sigma) at a final concentration in the medium of 1 μg/μl, for 2 days, before harvesting the cells in Trizol (Invitrogen) for total RNA preparation.

RNA Immunoprecipitation:

A HEK 293 cell line stably transfected with a tetracycline inducible FLAG tagged CYP33 plasmid (by Mark Koonce) was cultured in standard DMEM high glucose medium supplemented with 10% FBS (Gibco BRL) and 1% Pencillin/ Streptomycin (Gibco BRL) and 1μg/ml doxycycline. Peak induction of CYP33 was observed after 16

hours of induction (Mark Koonce; unpublished data). In order to precipitate endogenous transcripts with overexpressed CYP33, three dishes with 3.2 X 10^6 cells plated in 15 cm dish, grown for 2 days before Doxycycline (1 μg/ml) induction for 16 hours. After 16 h induction with Doxycycline, cells were washed with 1X PBS at room temperature. The cells were then trypsinized, washed with ice cold 1X PBS, counted and resuspend in 1X nuclei isolation buffer (NLB; see the composition below): 2 ml ice cold PBS + 2 ml NLB + 6 ml ice cold DEPC treated water. Nuclei were pelleted by centrifugation at 2,500G for 15 min and the nuclear pellet was resuspended in 2 ml ice cold Radio-Immunoprecipitation Assay (RIPA) buffer (containing the 1X protease inhibitor cocktail and 100 units/ml of RNAse inhibitor from Fermentas; see the composition below). 200 μl of this reaction was used for the isolation of total RNA (and 10% input) by Trizol (Invitrogen) method. The rest of the suspension was split as follows and the following steps were performed in the order: 1) 900 μl/ Dounce homogenizer using pestle B /control IgG pulldown/ and 2) 900 μl/ Dounce homogenizer; pestle B / Protein A agarose preclearing step/ AntiFLAG pulldown (Sigma).

The nuclei suspension is subjected to lysis using a Dounce homogenizer (15-20 strokes). Nuclear membrane and debris are pelleted by centrifugation at 13,000 rpm for 10 min. Preclearing of the samples was done using 10 μl of protein A agarose beads (Roche) and 10 μg of yeast tRNA (Sigma) at 4°C on SpeciMix Thermolyne rocker for 30min. Pull down steps involve: 1) 900 μl nuclear lysate, 5 μl of antiIgG (1 μg/μl), 20 μl of protein A agarose. Binding was done at 4°on SpeciMix Thermolyne rocker for 2h. 2) 900 μl nuclear lysate and 36 μl (total volume, not packed volume) of antiFLAG agarose (Sigma)

(10 µl of packed volume of the beads binds > 1 µg of FLAG fusion protein). After the binding step, the beads were washed for 5 times or more using NLB containing 1X protease inhibitor cocktail (from Roche). The RNA was eluted using 200 µl trizol per sample. cDNA was synthesized using the high capacity cDNA synthesis kit from Applied Biosystems (reaction scaled down to 10 µl). cDNA was diluted 1:2 (with 10 mM Tris pH 8) and 2 µl of it was used per 50 µl 1X Promega GoTaq Green mix PCR reaction. Buffer Compositions- NLB: 1.28 M sucrose; 40 mM Tris-HCl pH 7.5; 20 mM MgCl2; 4% Triton X-100 and RIP: 150 mM KCl, 25 mM Tris pH 7.4, 5 mM EDTA, 0.5 mM DTT, 1X mini protease inhibitor cocktail, 100 U/ml RNAsin (Fermentas).

Chromatin Immunoprecipitation:

ChIP was performed on 80% confluent MSA cells grown in 15 cm dishes, cross linked using 1% formaldehyde. The cells were sonicated (Branson Sonifier 250 sonicator with microtip) in lysis buffer supplied with the Millipore Magna ChIP A kit (Millipore) (1.4 X 10^6 cells/ 400 µl lysis buffer) under following conditions- output 4; 5X, 7sec each followed by output 5; 2X, 7 sec each) and harvested for ChIP using antibodies against MLL-N (Millipore; Anti MLL/HRX, N terminal, clone 4.4), MLL-C (Millipore; Anti MLL/HRX, C terminal, clone 9-12), CYP33 (Novus Biologicals, mouse polyclonal antiPPIE), H3K27Me3 (Millipore, rabbit polyclonal) and H3 (Millipore, anti H3). The protocol was followed as per (Millipore Magna ChIP manual). Semi-quantitative PCR was done on the precipitated DNA using primers listed in table (2).

Transinduction of *HOXC8* in MSA cell line:

MSA cells were maintained at a cell density of 70-80% in DMEM F-12 (50%-50%). The cells were plated at a density of 400,000 cells/well in a six well plate for transfection with individual constructs made from the parental plasmid pCMV-HA and intergenic transcripts *NC3* and *NC4*. The cDNA for *NC3* and *NC4* were subcloned from pGEMTEasy (Promega) into pCMVHA (Clontech) at EcoRI/NotI and SalI/BstZI sites respectively. To generate promoter-less control constructs containing *NC3* and *NC4* cDNA, each cDNA was cloned into pCMVHA at SphI and SpeI after removing the promoter containing fragment. The controls for the experiment included the same constructs without the promoter. The cells were transfected 2h with the above plasmids after plating them in a six well plate, using 4 μg of each plasmid construct and 10 μl Lipofectamine 2000 reagent (Invitrogen) following the manufacturer's protocol. The cells were incubated at 37°C for 48 hours without intervening media change. The cells were harvested after 48 hours for isolation of total RNA using Trizol (Invitrogen). First strand cDNA was synthesized using 4 μg (or more) of the total RNA, random hexamers and reagents from Applied Biosystems High Capacity Reverse Transcription Kit. The cDNA was then used for quantitative PCR using Promega GoTaq qPCR Master Mix. This experiment was repeated thrice and results evaluated for fold change in *HOXC8* expression using t test.

Table 2: Primers and probes used in this study.

Primer/ Probe (Experiment)	Sequence
hHOXC8-F1 (RT-PCR; MSA treated with CsA)	5’ TAGCTGCCACGGAGACGCCTC 3’
hHOXC8-R1 (RT-PCR; MSA treated with CsA)	5’ AACGAAACTTCAAGGGAGTTGC 3’
hHOXC8-F2 (RT-PCR; MSa treated with CsA)	5’ ATCCTTATTTGACACGAAAACGTC 3’
hHOXC8-R2 (RT-PCR; MSA treated with MSA)	5’ AAATAAAGAGTGGGGGAAGTCC 3’
FLAG-CYP33-F (RT-PCR; RNA-IP for ensuring overexpression)	5’ ATGGCTGACTACAAGGACG 3’
FLAG-CYP33-R (RT-PCR; RNA-IP for ensuring overexpression)	5’ CAACCAGTCATCATCTGACC 3’
hGAPDH-F (RT-PCR; MSA treated with CsA)	5’ GAAGGTGAAGGTCGGAGTC 3’
hGAPDH-R (RT-PCR; MSA treated with CsA)	5’GAAGATGGTGATGGGATTTC 3’
hNC1-F (Mark Koonce) (RT-PCR in RNA-IP, Quant. PCR in NC1 upstream transcription)	5’GGCAACGGCACAGAATGAGG 3’
hNC1-R (Mark Koonce) (RT-PCR in RNA-IP, Quant. PCR in NC1 upstream transcription)	5’ GTTTCCCTGGCAATGGTTAG 3’
hF3 (Quant. PCR in NC1 upstream transcription)	5’ AGGTGGGTATGGGCAATCTGA 3’
hR3 (Quant. PCR in NC1 upstream transcription)	5’ ATCCCCGGGGTCCCTTTCTTAG 3’
hF2 (Quant. PCR in NC1 upstream transcription)	5’ GGGCCAGCGCTTCCTTGTCGTA 3’
hR2 (Quant. PCR in NC1 upstream transcription)	5’ AGGGGTTGCAGGCAGGTCACTCTC 3’
hF1 (Quant. PCR in NC1 upstream transcription)	5’ TGCTCCCGCCGTCCCATTA 3’
hR1 (Quant. PCR in NC1 upstream transcription)	5’ GCATTATTCCAGCTCTACGGGTTGA 3’
hF4 (Quant. PCR in NC1 upstream transcription)	5’CCCTCATCTGTCTAGCTCCTGGTC 3’
hR6 (Quant. PCR in NC1 upstream transcription)	5’AGCCTTTCTCGGACTGCAGG 3’
hF6 (Quant. PCR in NC1 upstream transcription)	5’ AAGTGAGCTCGTAGGCAACCAG 3’
hR7 (Quant. PCR in NC1 upstream transcription)	5’ AGCTGACCAATTGCCAAGGTG 3’

hLE1a-F (or NC1spliceF) (RT-PCR with hHOXC6-R, expression in 293 cell line for *NC1-HOXC6* variants)	5’ GGCAGCGGCACAGAATGAGG 3’
hHOXC6-R (RT-PCR, expression in 293 cell line for *NC1-HOXC6* variants)	5’ TGTCGCTCGGTCAGGCAAAG 3’
hHOXC6-F (RT-PCR, expression in 293 cell line for *HOXC6*)	5’ CACAGACCTCCATCGCTCAGGAT 3’
hAS-F (RT-PCR, expression in 293 cell line for NC2)	5’ AATGACTGCCTGCTGGCT 3’
hAS-R (RT-PCR, expression in 293 cell line for NC2)	5’ AGGGGTTGCAGGCAGGTCACTCTC 3’
hHc8LE-F (RT-PCR, RNA-IP for NC2 pull down)	5’ CAACCGACCCTCAGTGC 3’
hHc8LE-R (RT-PCR, RNA-IP for NC2 pull down)	5’ TTTCTCTTCCCTCATTCTGTG 3’
hLTCloning-F (Cloning, RT-PCR with hLTR1 for expression in 293 cell line for NC4)	5’ GCAAGCAGAGAAGGCATAGCAG 3’
hLTR1 (RT-PCR for expression in 293 cell line for NC4)	5’ ATTCTCCTTAGCTAGGAACCAGC 3’
mhCYP33-F (RT-PCR, mES differentiation)	5’TCGTGTCAATTTGGCCAAGC 3’
mhCYP33-R (RT-PCR, mES differentiation)	5’TGCAGTGGGCTCTCCCTCCTG 3’
MLL-PHDmidF (RT-PCR)	5’ CTAATCTGCCAGAAAGTGTGG 3’
MLL-PHDmidR (RT-PCR)	5’ GACCTTTGGGTTTGGGAGCTGG 3’
HsHc8LE2pcrF (RT-PCR, expression in 293 cell line for NC3)	5’ GTAGGACTTGTGTGTCG 3’
HsHc8LE2pcrR (RT-PCR, expression in 293 cell line for NC3)	5’CTCACCGGTCGGCGATTC 3’
hMLLRltF (Quant. PCR; MSA cell line)	5’ GCATCGATGACAACCGACAGTG 3’
hMLLRltR (Quant. PCR; MSA cell line)	5’ CACAGCCATATGCACATTCTTC 3’
hCYP33RltF (Quant. PCR; MSA cell line)	5’ GCAGCAGCTATCGACAACATGAATG 3’

hCYP33RltR (Quant. PCR; MSA cell line)	5’ CTTCAACCAGTCATCATCTGACCAG 3’
hHOXC8RltF (Quant. PCR; MSA cell line)	5’ CTTAAATCAAAACTCGTCTCCCAGC 3’
hHOXC8RltR (Quant. PCR; MSA cell line)	5’ TAGTTCCAAGGTCTGATACCGGC 3’
mLE1bF (RT-PCR, with hHc6R for expression in MEF of *NC1-Hoxc6* variants)	5’ATAGGACTGGTGTGTAGGCAGAGG 3’
mHc6R (RT-PCR)	5’TGTCGCTCGGTCAGGCACAG 3’
mHc6pF (or mHoxc6F) (RT-PCR, with hHc6R for expression in MEF of *Hoxc6*)	5’AAAGAAATCATAGCCCGACCAGG 3’
mLT-F (RT-PCR; expression in MEF of NC4)	5’TTGTGTTCTGAAATCAGGCTTGG
mLT-R (RT-PCR; expression in MEF of NC4)	5’CCTCTCTGATTCTACACCTGGCAG 3’
mhAS-F (RT-PCR; expression in MEF of NC2)	5’TTTTTTTTTTTTTATTCTAATCTC 3’
mhAS-R (RT-PCR; expression in MEF of NC2)	5’AAAAAGTTCACGTTCATGGATGG 3’
mHoxc8-F (RT-PCR; MEF)	5’CCTATTACGACTGCCGGTTCC 3’
mHoxc8-R (RT-PCR; MEF)	5’CCACTTCATCCTCGATTCTGG 3’
mActin-F (RT-PCR; MEF, mES differentiation)	5’GCCAGGTCATCACTATTGGCAACGAG 3’
mActin-R (RT-PCR; MEF, mES differentiation)	5’GCCACCGATCCACACAGAGTACTTG 3’
mLTFa (RT-PCR; expression of NC4, mES differentiation)	5’CCTGTATTTTGTCGGACCCTGTC 3’
mLTRd (RT-PCR; expression of NC4, mES differentiation)	5’ CCTCTCTGATTCTACACCTGGCAG 3’
mBrachyuryF (RT-PCR; mES differentiation)	5’GACTTCGTGACGGCTGACAA 3’
mBrachyuryR (RT-PCR; mES differentiation)	5’CGATGTGAATCCGAGGTTC 3’
mFlk-1F (RT-PCR; mES differentiation)	5’CACCCCAGATCGGTGAGAAA 3’
mFlk-1R (RT-PCR; mES differentiation)	5’AATCAGGGCATATTGGTTTTTGG 3’

mF4 (RT-PCR; mES differentiation for NC2)	5'AAGTATTTAAAATTCGGGACAGC 3'
mLE1aGSPR (RT-PCR; mES differentiation for NC2)	5' CATTCTGTGCCGTTGCCGAG 3'
mF3 (RT-PCR)	5'CTCTTATGGGGCCTGCAAAC 3'
mR3 (RT-PCR)	5'ACGGAGAGCTGGACCAGGATT 3'
mHoxc8RltF (Quant. PCR; mES differentiaiton)	5' GACGCCTCCAAATTCTATGGC 3'
mHoxc8RltR (Quant. PCR; mES differentiation)	5' AGACGAGTTCTGATTTAAGTGGCC 3'
mCyp33RltF (Quant. PCR; mES differentiation)	5' GCAGCAGCTATCGACAACATGAATG 3'
mCyp33RltR (Quant. PCR; mES differentiation)	5' CTTCAACCAGTCATCATCTGACCAG 3'
hU1F (RT-PCR; RNA-IP)	5' ATACTTACCTGGCAGGGGAG 3'
hU1R (RT-PCR; RNA-IP)	5' CAGGGGGAAAGCGCGAACGCA 3'
hHOXA9-F (RT-PCR; RNA-IP)	5' CACTTTGTCCCTGACTGCCTATGC 3'
hHOXA9-R (RT-PCR; RNA-IP)	5' GGTACATGTTGAACAGAAACTCTTTCTC 3'
mMLLRltF (Quant. PCR; mES differentiation)	5' GCATCGATGACAACCGACAGTG 3'
mMLLRltR (Quant. PCR; mES differentiation)	5' CACAGCCATATGCACATTCTTC 3'
mCyp33RltF (Quant. PCR; mES differentiation)	5' GCAGCAGCTATCGACAACATGAATG 3'
mCyp33RltR (Quant. PCR; mES differentiation)	5' CTTCAACCAGTCATCATCTGACCAG 3'
hHOXC8promoterF (PCR, ChIP)	5' ACCTACCGACAGTGAGGAGCG 3'
hHOXC8promoterR (PCR, ChIP)	5' TCTGGCTCACGAGTACCCCG 3'
hHOXC8exon1F (PCR, ChIP)	5' CCCCCAGCCCTGTACCCCTCAACC 3'
hHOXC8exon1R (PCR, ChIP)	5' CGGCGGCGCTCCTCACTGTCG 3'
hHOXC6promoterF (PCR, ChIP)	5' GCCCATTTGCCCACTCC 3'

hHOXC6promoterR (PCR, ChIP)	5’ GCCCCCTTCTAAAATTACATCCTG 3’
hGAPDHpromoterF (PCR, ChIP)	5’ TACTAGCGGTTTTACGGGCGCACGT 3’
hGAPDHpromoterR (PCR, ChIP)	5’ TCGAACAGGAGGAGCAGAGAGCGAA 3’
hCyp33CeltaRRM-BamH1-F (Cloning)	5’ CTCGAGGCCTCACACGTACTC 3’
hCyp33-Xho1-R (Cloning)	5’ CTCGAGGCCTCACACGTACTC 3’
hCyp33-BamH1-F (Cloning)	5’ GGATCCGCCACCACCAAGC 3’
hCyp33DeltaRRM-BamH1-F (Cloning)	5’ GGATCCGGCTCTTCCAGGCCAG 3’
hHc8LE1Sal1F (Cloning NC1 no promoter control via pGEMTEasy; transinduction)	5’GTCGACCCGACCCTCAGTGC 3’
hHc8LE1Sal1F (Cloning NC1 no promoter control via pGEMTEasy; transinduction)	5’GTCGACTCCCTCATTCTGTG 3’
hHc8LE2Sal1F (Cloning NC3 no promoter control via pGEMTEasy; transinduction)	5’GTCGACAGCCATCATAAAGG 3’
hHc8LE2Sal1R (Cloning; NC3 no promoter control via pGEMTEasy; transinduction)	5’GTCGACCCTCAGGATGGTGC 3’
hLTSal1F (Cloning NC4 no promoter control via pGEMTEasy; transinduction)	5’ ACGCGTCGACCAAGACAGAGGCAAG 3’
hLTSal1R (Cloning NC4 no promoter control; transinduction)	5’ ACGCGTCGACTCCTTTCATGTTTTTTC 3’
hLTcloningHindIIIF (Cloning NC4 via pGEMTEasy; transinduction)	5’ CCCAAGCTTCAAGACAGAGGCA 3’
hLTcloningXhoIR (Cloning NC4 via pGEMTEasy; transinduction)	5’ CCGCTCGAGTCCTTTCATGTTTTT 3’
RltHPRT1F (Quant. PCR; transinduction)	5’ TGACACTGGCAAAACAATGCA 3’
RltHPRT1R (Quant. PCR; transinduction)	5’ GGTCCTTTTCACCAGCAAGCT 3’
HsHc8Cloning-F (RT-PCR, Cloning)	5’ CGCAGCCATCATAAAGGCCTC 3’
hLTR1 (RT-PCR, Cloning)	5’ ATTCTCCTTAGCTAGGAACCTGC 3’
HsHc8PCR-F (RT-PCR)	Same as Mark Koonce hNC1-F

HsHc8PCR-R (RT-PCR)	Same as Mark Koonce hNC1-R
HsHc8LE2pcrF (RT-PCR; NC3 expression in human)	5’ AGTAGGACTTGTGTGTCGGCAGAG 3’
HsHc8LE2pcrR (RT-PCR; NC3 expression in human)	5’ CTCACCGGTCGGCGATTC 3’
hLTCloning (RT-PCR; NC4 promoter in ChIP)	5’ GCAAGCAGAGAAGGCATAAGCAG 3’
hLTR1 (RT-PCR; NC4 promoter in ChIP)	5’ ATTCTCCTTAGCTAGGAACCAGC 3’
mirLet7c (Northern Blot)	5’ AACCATACAACCTACTACCTCA 3’
misLet7e (Northern Blot)	5’ ACTATACAACCTCCTACCTCA 3’
hCypRTEcoF (Cloning)	5’GAATTCGCGCGCGAGCAAGATGG
hCypRTNotR (Cloning)	5’GCGGCCGCCACGTACTCCCCACA
AAUU containing oligonucleotide (or SlxCypF) (REMSA after T7 pol. based in vitro transcription)	5’ ATTCGTTAATACGACTCACTATAGGAACGCAATTATTCGTATTTAGAACGCAATTATTCGTATTTAGAACGCAATTATTCGT 3’
AAUU containing oligonucleotide (or SlxCypR) (REMSA after T7 pol. based in vitro transcription)	5’ ACGAATAATTGCGTTCTAAATACGAATAATTGCGTTCTAAATACGAATAATTGCGTTCCTATAGTGAGTCGTATTAACGAAT 3’
Control (or SlxRRM-F) (REMSA after T7 pol. based in vitro transcription)	5’ ATTCGTTAATACGACTCACTATAGGCATACTCCCTTAGTCACATACTCCCTTAGTCACATACTCCCTT 3’
Control (or SlxRRM-R) (REMSA after T7 pol. based in vitro transcription)	5’ AAGGGAGTATGTGACTAAGGGAGTATGTGACTAAGGGAGTATGCCTATAGTGAGTCGTATTAACGAAT 3’

CHAPTER FOUR

RESULTS

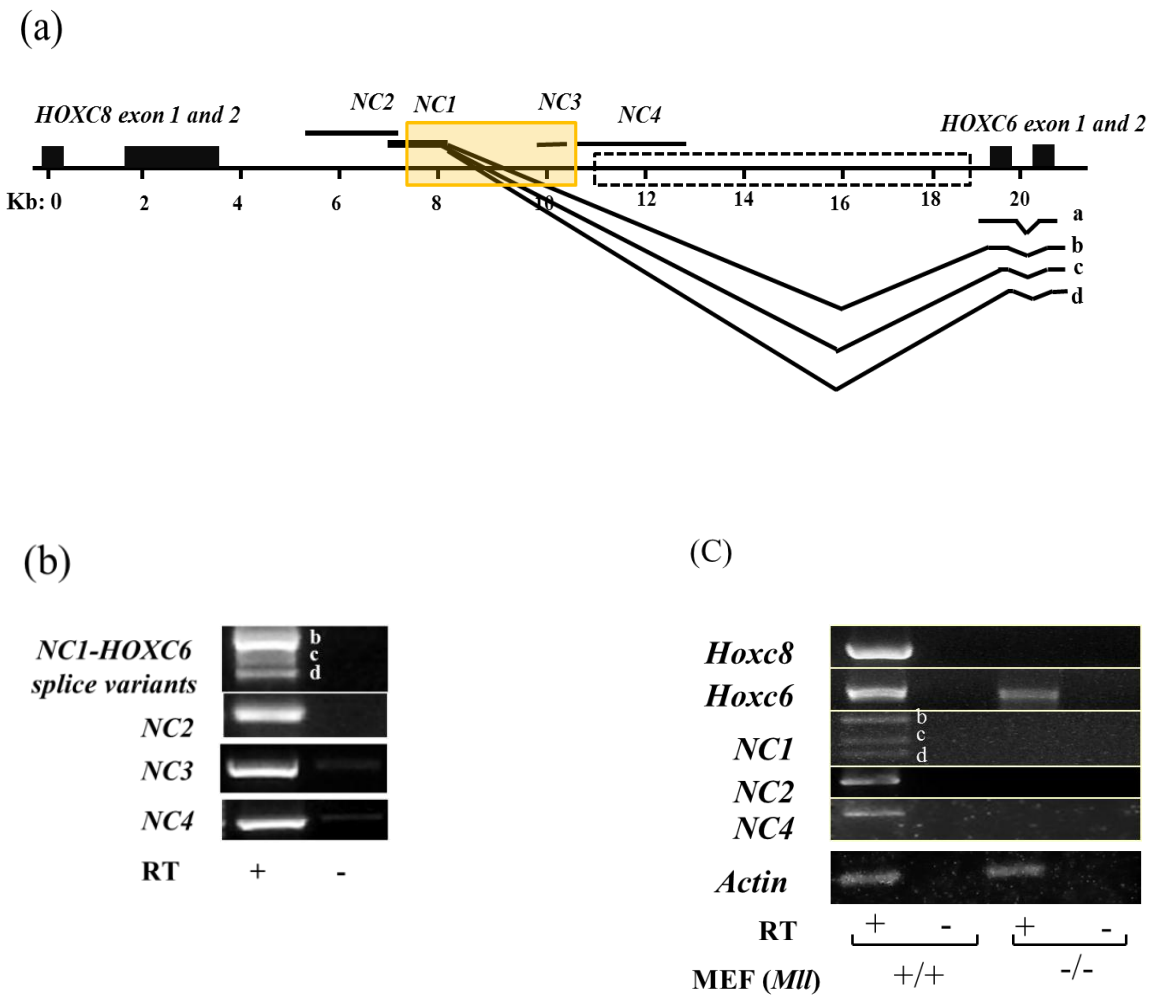
The *Hoxc8- Hoxc6* intergenic region contains the *Hoxc8 3'* regulatory region and four major ESTs that are conserved and expressed in mouse and human:

The 15 kb *Hoxc8-Hoxc6* intergenic region contains a 8 kb *Hoxc8 3'RR* (Bradshaw et al. 1996, Deschamps et al. 1999) required for the maintenance of *Hoxc8* expression during mouse embryogenesis; a process that also entails role of MLL (Hanson et al. 1999). This study aimed first at characterizing the *Hoxc8-Hoxc6* intergenic region in terms of its inter-species conservation, presence of ESTs, expression in human and mouse cell lines, MLL dependency, promoter prediction and splice site conservation. In terms of conservation of the *HOXC8-HOXC6* intergenic region, when compared to *Homo sapiens*, some mammalian species show more than 80% identity (mouse- 87%, chimpanzee- 98%, cow- 82% and rat- 84%). A cluster of YY1 binding sites (core consensus of CCAT) was found present just outside of the 5' end of the 3' regulatory region. YY1 is a human homologue of *Drosophila* Pleiohomeotic (Pho) that functions as a Polycomb group protein mediating gene silencing via its binding to DNA GCCAT consensus sites located in the *Drosophila* PREs (Ringrose and Paro 2007) . PREs have not been fully

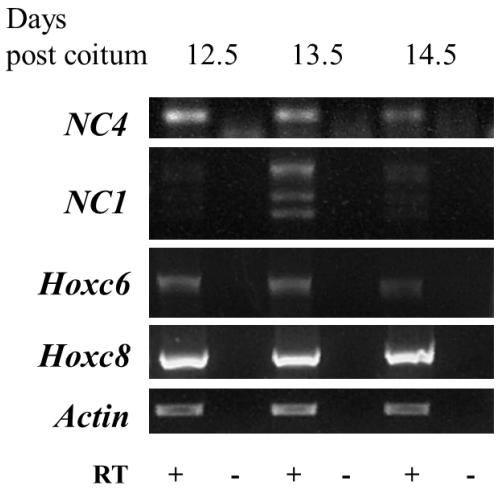
characterized in mammals yet. YY1 binding sites; however, may be part of PREs in mammals since YY1 is a human homologue of Drosophila Polycomb group protein called Pleiohomeotic. YY1 has been proposed to mediate enhancer-promoter interactions on the imprinted loci (Kim et al. 2006). As mentioned before, early during embryogenesis (cellular blastoderm stage) in *Drosophila*, the TRE/PREs show co-localization of both the PcG and TrxG groups of proteins on these regulatory elements (Orlando et al. 1998). Since CYP33 functions as a Polycomb group protein in *Drosophila* (Andrew Dingwall; unpublished data) and recruits repressive proteins to the MLL complex, it is important to determine if the *HOXC8* enhancer region contains binding sites for PcG proteins like YY1. Importantly, four major transcripts identified as ESTs were identified in this region. These were named *NC1* (for Non-coding 1), *NC2*, *NC3* and *NC4* [Fig. 9a]. *NC1* is a transcript highly conserved amongst diverse animal species and transcribed from the same strand as *HOXC8. NC1* has been shown in human cells to splice into exon 1 of *HOXC6*, *HOXC5* and *HOXC4* (Boncinelli et al. 1989). *NC2*, on the other hand, is transcribed from the complementary strand and its transcription unit partially overlaps with *NC1* [Fig. 9a]. EST data entry for *NC2* suggests that it is polyadenylated (Table 3). *NC3* is an EST found 5' to the *Hoxc8 3'*RR. It splices with the downstream *Hoxc6* in mouse (Table 3). *NC4* is the longest identified polyadenylated transcript (2.8 kb) in this region, without any splicing potential. In order to determine if the four transcripts mentioned here are expressed in mouse and human, RT-PCR analysis was done using RNA from mouse embryonic fibroblasts and the human HEK293 cell line.

Figure 9: The *Hoxc8-Hoxc6* region encompasses a regulatory region important for maintenance of *Hoxc8* expression. (a) The *Hoxc8-Hoxc6* intergenic region spans15 kb with a previously identified 8 kb 3' regulatory region (*Hoxc8 3'* RR) required for *Mll* dependent maintenance of *Hoxc8* expression during mouse embryogenesis shown by dotted rectangle. Four ESTs were mapped in this region with a highly conserved *NC1* transcript about 7 kb downstream of the *Hoxc8* TSS that splices into three splice-acceptor sites within *Hoxc6 exon1* depicted here as transcripts b, c and d. Transcript a is the canonical *Hoxc6* transcript. *NC2* is a transcript from the complementary strand that is polyadenylated and partially overlaps *NC1*. *NC3* is an EST mapped just 5' to the *Hoxc8 3'* RR and can potentially splice into *Hoxc6 exon1*. Importantly a 2.8 kb long *NC4* shows no splicing potential and overlaps almost completely with the *Hoxc8 3'* RR. Shaded yellow rectangle shows the chromatin region with bivalent histone modifications found in the mouse embryonic stem cells. The ESTs are shown as lines above the main transcription unit whereas the protein coding gene exons are depicted as black boxes. (b) Semi-quantitative RT-PCR analysis for gene expressions in 293 HEK cell line shows transcription from the *NC1*, *NC2*, *NC3* and *NC4* regions. Three different *NC1-HOXC6* splice variants were detected in this cell line. (c) Semi-quantitative RT-PCR analysis in mouse embryonic fibroblasts (MEF) shows expression of protein coding transcripts such as *Hoxc8* and *Hoxc6* as well as *NC1*, *NC2* and *NC4*. Three splice variant forms of *NC1-Hoxc6* were detected in this cell line. *NC3-Hoxc6* splice variant has already been reported as EST (see table 3). When compared between wild type and *Mll* null MEF, the expression of *Hoxc8* and the intergenic transcripts was found to be down regulated in *Mll*

null MEF. (d) Semi-quantitative RT-PCR analysis in mouse embryos from days 12.5, 13.5 and 14.5 days post coitum. Expression of *Hoxc8* and *Hoxc6* is observed in all the three stages of mouse embryos tested. Expression of *NC4* was found expressed at higher levels in day 12.5 embryo with its expression declining in the later stages. All three forms of *NC1-Hoxc6* splice variants are detectable in the day 13.5 mouse embryo and not in 12.5 and 13.5 days old embryos suggesting its regulated expression.



(d)



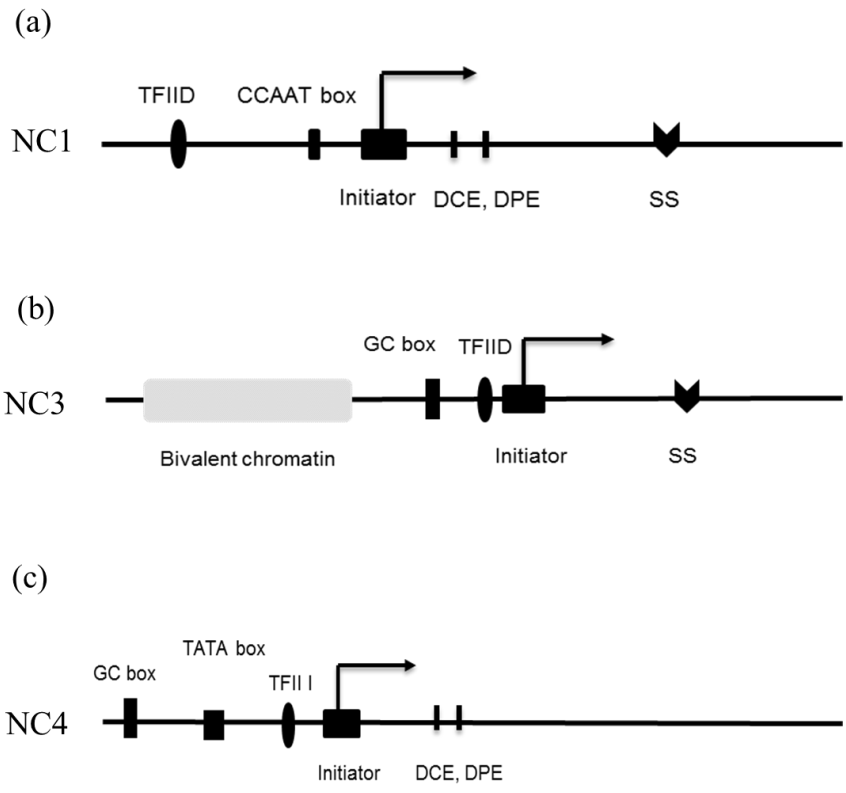
Using RT-PCR analysis of mouse embryonic fibroblasts (MEFs) RNA, *NC1* was found to splice to three different acceptor sites within exon 1 of *Hoxc6* [Fig.9a, c and d]. In order to test if the two novel splice isoforms of *NC1-Hoxc6* identified in MEFs were not an artifact of in vitro cultured MEFs, their expression was tested in late stage mouse embryos [Fig. 9d]. All the three isoforms of *NC1-Hoxc6* were identified in 13.5 dpc embryos with only very weak expression seen on days 12.5 and 14.5 [Fig. 9d]. This is indicative of endogenous expression of *NC1* splice isoforms during development. All the intergenic transcripts from the *Hoxc8-Hoxc6* region show relatively low abundance (< 0.3% as compared to the abundance of endogenous actin; quantitative RT-PCR data not shown). By comparing RNA from *MLL* wild type and *MLL* null MEFs the expression of *Hoxc8* as well as that of intergenic transcripts were found to be *MLL* dependent [Fig. 9c]. *Hoxc6* expression is slightly down regulated in the absence of MLL.

Table 3: Protein coding potential of intergenic transcripts from *HOXC8-HOXC6* region in human and mouse.

Human					
Name	Accession number	Length (nucleotides)	Splicing with	Predicted ORFs	CPC score
NC1-HOXC6 variant 1	AK314829	2047	*HOXC6 exon1*	+; HD (Homeo domain)	-0.853936
NC1-HOXC6 variant 2	This study	1929	*HOXC6 exon1*	+; HD	-0.48368
NC1-HOXC6 variant 3	This study	1864	*HOXC6 exon1*	+; HD	-0.846765
NC2 (AS)	BC020349	1622	Not reported	–	-1.13373
NC3-HOXC6	This study	1823	*HOXC6 exon1*	+; HD	-0.789837
NC4	BC041835	2811	None	–	-1.1123
Mouse					
NC1-Hoxc6 variant 1	3720901	2140	*Hoxc6 exon1*	+; HD	-0.988327
NC1-Hoxc6 variant 2	This study	2020	*Hoxc6 exon1*	+; HD	-0.823132
NC1-Hoxc6 variant 3	This study	1957	*Hoxc6 exon1*	+; HD	-0.819038
NC2	BC125340	1890	Not reported	–	-0.604935
NC3	BI412975.1	1572	*Hoxc6 exon1*	+; HD	-0.438619
NC4	This study	2565	None	–	-0.931389

Upstream regions of *NC1*, *NC3* and *NC4* were tested for the presence of promoter specific features such as the TFII D/I binding sites, TATA/CCAAT or CpG boxes as well as the transcription initiator elements. The putative transcriptional start site of all the intergenic transcripts show the presence of TFIID factor consensus binding sites in the promoter proximal region implying the role of Polymerase II in transcription [Fig. 10]. Furthermore, *NC1* also has a 5' donor splice site that splices to three different 3' acceptor splice sites in the first exon of *HOXC6*. The *NC4* transcription unit also has DPE (downstream promoter element) and DCE (downstream core element) consensus sequence elements just downstream of the transcription start site. The presence of promoter specific features at these transcriptional units suggests the presence of independent promoters and regulation of the intergenic transcription.

Figure 10: The *HOXC8-HOXC6* region intergenic transcripts are transcribed by RNA Polymerase II. (a) *NC1*, (b) *NC3* and (c) *NC4* predicted promoter proximal elements. The CCAAT or TATA boxes are indicated by black rectangles. The initiator element is shown here in black box that contains the putative transcription start site (arrow). The ovals denote the TFIID/I binding sites indicating a RNA Polymerase II binding sites. The predicted splice sites are indicated by chevron (SS). DPE: downstream promoter element and DCE: downstream core element. Diagram is not drawn to scale.



The protein coding potential of the *HOXC8-HOXC6* intergenic transcripts was calculated using the Coding Potential Calculator (CPC) (Kong et al. 2007), NCBI ORF finder, conservation of sequence, conservation of position in the locus, presence of promoter proximal elements, conservation of exon-intron architecture, a valid open reading frame (of more than 300 nts), potential splicing sites and polyadenylation signals. The results are summarized in Table 3. The CPC is based on the design of a Support Vector Machine (SVM) that takes into account six criteria used to deduce the longest Open Reading Frame (ORF), estimate evolutionary conservation based log odds score denoting its quality and compare protein sequence from the predicted ORFs with UniProt database. The SVM scores thus derived represent the confidence level of the prediction. The farther away the scores are from zero; higher is the reliability of the prediction. The negative scores represent prediction of non-coding potential. *NC1* splicing with the downstream *HOXC6, HOXC5* and *HOXC4* has been reported for human embryos (Zeleznik-Le, Harden, and Rowley 1994). The current study found *NC1* splicing at three distinct splice sites within the *HOXC6 exon1* thereby giving rise to three splice variants both in human and mouse [Fig. 9b, c and d]. All of the three splice variants are capable of producing a homeo domain containing protein. In addition to the widely accepted threshold for valid ORFs of 100 amino acids (Dinger et al. 2008b), the presence of a valid Kozak sequence close to the predicted start codon was found only in the *NC1-HOXC6* splice variant 1, from both mouse and human, which is also found to be translated in mouse and *Xenopus* (Oliver et al. 1988, Bardine et al. 2009). The *NC1-HOXC6* splice variants 2 and 3 however, lack a valid Kozak sequence in front of the first start codon in

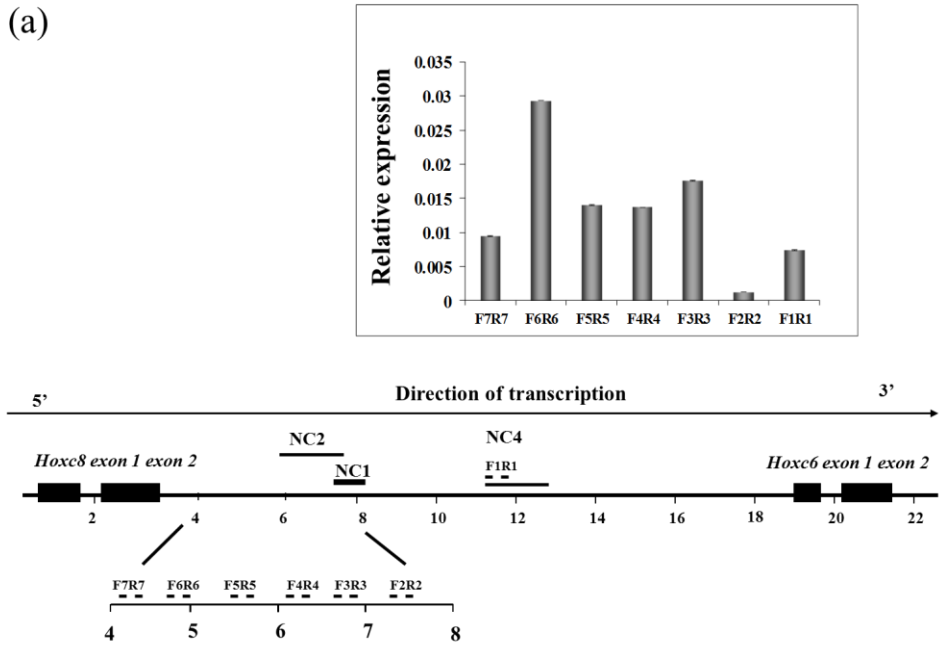
all the species examined (human, mouse, rat, cow and chimpanzee) and the hypothetical protein sequence produced is not conserved between species weakening the likelihood of their expression in vivo. The *NC1-HOXC6* splice variant 1 may, therefore, not be a true non-coding transcript and instead represent an alternative untranslated first exon of *HOXC6* similar to the alternative non coding first exon of mouse *HOXA9* (Erfurth et al. 2008). All the three sense reading frames and none of the amino acids sequences thus formed is found conserved between species thereby lessening the possibility of these splice variants having any functional significance as a protein coding genes (Dinger et al. 2008a). Likewise, *NC3* splices with *HOXC6* [Fig. 9b] potentially yielding a homeodomain containing protein; however, neither is there a valid Kozak sequence present in the ORF nor is the protein sequence thus obtained found conserved in other species thereby minimizing the likelihood of its translation into protein. *NC4* is a bonafide non-coding transcript that does not show splicing and has no long open reading frame. Even though *NC1* and *NC3* are found spliced with *HOX* genes, they may have a regulatory role for the *HOXC6* transcription unit. *NC1*, in particular, is an interesting transcript since it has six nucleotides or longer AT tracts that may serve as potential binding sites for CYP33. Pertaining to the expressions of *HOXC8-HOXC6* intergenic transcripts from both mouse and human, the following was also experimentally determined: the whole region (in addition to the EST containing region) between the *HOXC8* 3' UTR and *NC1* was tested for the occurrence of transcription [Fig. 11a]. The existence of a miRNA arising from the region that shows ultra-conservation between

diverse species in mouse embryonic fibroblasts was investigated [Fig. 11b] but we could not find evidence of such miRNA in the MEFs tested.

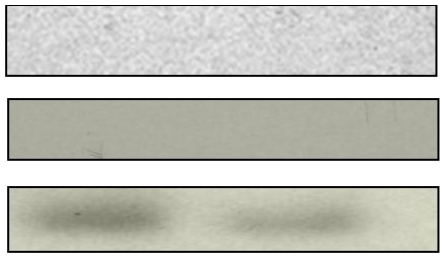
Figure 11: Expressions of intergenic transcripts *NC1, NC2, NC3* and *NC4* in human and mouse.

(a) Real time quantitative RT-PCR performed on total RNA from 293 cell line to look for transcription in the region of upstream of *NC1*. As shown, the entire region between 4 kb to 8 kb (from the *Hoxc8* TSS) is transcribed at different levels. The primers used to amplify the different regions are shown as pairs next to the region under investigation. The expression from the region between 5.8 kb and 6.8 kb may represent transcription from either or both the strands (*NC1* and *NC2* overlap). (b) Northern Blot to test for the presence of miRNA in the ultra-conserved *NC1* region in mouse embryonic fibroblasts. Probe for *NC1* miRNA (both forward and reverse strand) was designed to the region conserved between diverse species. Mouse microRNA let7c and let7e were used as both positive control and for normalization.

(a)



(b)



NC1 R

NC1 F

Mouse mirlet7c + mirlet7e

MEF ***Mll*** **WT** ***Mll*** **null**

Non-coding transcript *NC4* is transcribed earlier and in greater amounts than *Hoxc8* in differentiating mouse embryonic stem cells.

Regulatory elements of the clustered gene such as enhancers, in many instances, are transcribed into potentially non-coding RNA (Ho et al. 2009, Lempradl and Ringrose 2008). The best studied example in the literature is that of certain iab (infra abdominal) elements e.g. iab5 that supports the expression of the *Abdominal B* (*AbdB*) gene in segment 10 of *Drosophila*. Studies have discovered a correlative expression of these elements with that of *AbdB* suggesting a possible regulatory role. Similar studies are lacking for mammalian *Hox* genes that are present in clusters with their cell type specific enhancers present in the intergenic regions (Juan and Ruddle 2003, Bradshaw et al. 1996, Tschopp et al. 2009). In order to determine if the conserved *NC4* ncRNA has a role in *Hoxc8* regulation, it becomes essential to compare the expression kinetics of *NC4* as well as *Hoxc8* during development. In vitro differentiation of mouse embryonic stem cells (mES) into numerous lineages serves as an ideal in vitro model of studying cellular processes occurring during development since it at least partially mimics embryogenesis (Wobus and Boheler 2005, Ling and Neben 1997). I performed a one-step differentiation of RW4 mouse embryonic stem cells in vitro by LIF removal to obtain Embryoid Bodies (EBs) that were harvested each day up to 10 days to harvest total RNA (Ling and Neben 1997, Schmitt, Bruyns, and Snodgrass 1991, Lu et al. 2002). *Hoxc8* expression has been noticed both in developing neural tube (Shashikant and Ruddle 1996) as well as in hematopoietic cells of varied lineages (Argiropoulos and Humphries 2007; Shimamoto et al. 1999; Magli, Largman, and Lawrence 1997). In order to ensure that the induced mES

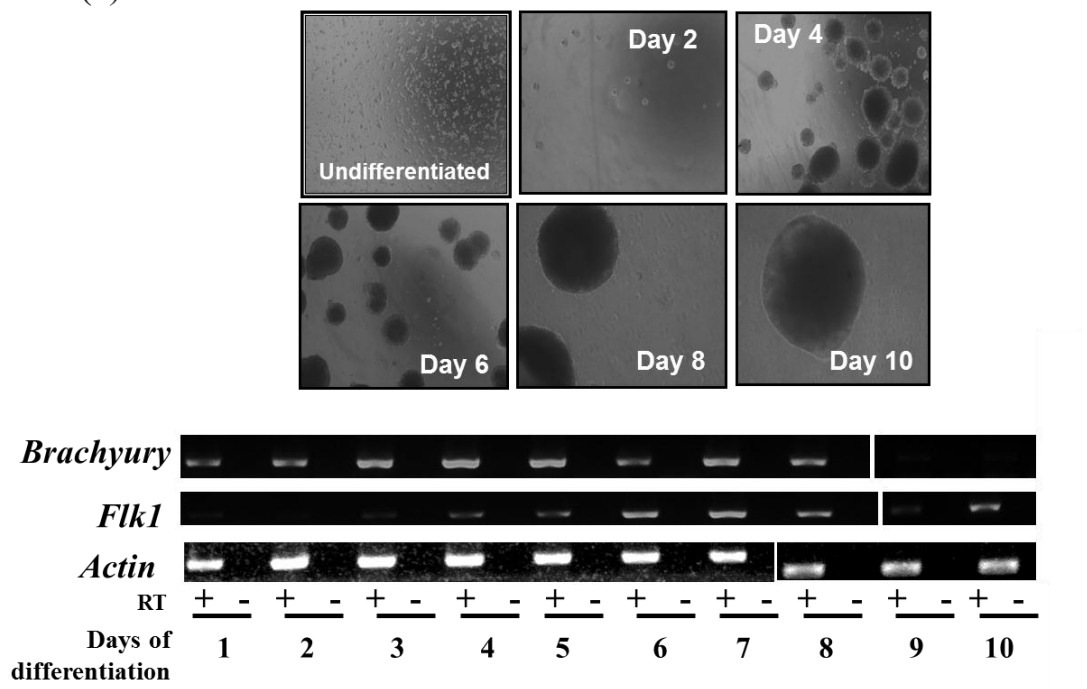
cells were differentiating into cells of hematopoietic lineages too, the course of differentiation was monitored by the expressions of Flk-1 (marker for endothelial differentiation) and Brachyury (for mesodermal differentiation) (Leahy et al. 1999) [Fig. 12a]. Using both semi-quantitative and quantitative real time RT-PCR approaches, I determined the expression profiles of *NC4* and *Hoxc8* during EB differentiation. *NC4* expression begins by day 1 of LIF removal indicating an early induction and decondensation of the chromatin at the 3' regulatory region of *Hoxc8* that expresses *NC4*. The expression of *NC4* then gradually increases through the 10 day-differentiation period [Fig. 12b]. *Hoxc8* expression shows a small burst of expression beginning by day 2 and declining by day 4 [Fig. 12b and d]. Compared to *NC4*, however, the expression levels of *Hoxc8* are low. *Hoxc8* expression reappears by day 6 and steadily increases up to day 10. An overall comparison of the ncRNA *NC4* and the protein coding transcript *Hoxc8* indicates that the *NC4* expression precedes that of *Hoxc8*. Additionally, post day 6 as seen by quantitative RT-PCR analysis, the expression of *NC4* is found about 1.25 to 2.5 fold in excess of the *Hoxc8* expression [Fig. 12d]. The expression of *Mll* and *CYP33* transcripts was monitored in WT mES cells and as noted in Fig. 12e, *CYP33* levels show a 2.8 fold increase by day 3 of differentiation as compared to the basal level expression at day 1; however the levels decrease by day 6, which roughly corresponds to the time during which the second phase of *Hoxc8* expression begins. A similar analysis for *Mll* expression in WT EBs showed a 2.5 fold increase in the levels by day 4 of differentiation and a subsequent decrease to basal level by day 7. *Mll* transcript expression post day 7 of

differentiation appears to show a trend of a steady increase in the expression concomitant with the increase in *Hoxc8* during its second phase of expression.

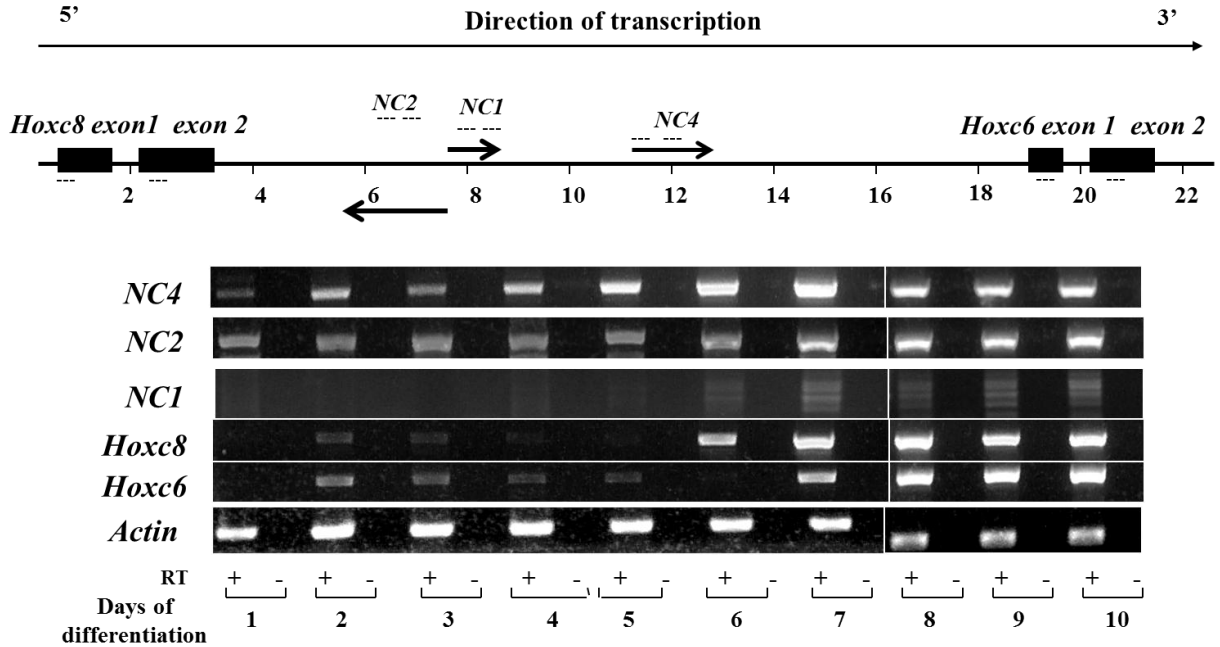
Figure 12: Expression patterns of *Hoxc8* and intergenic transcripts in *in vitro* differentiating EBs.

(a) Light micrographs of *in vitro* differentiating RW4 mouse embryoid bodies (EB) from different days of differentiation. Differentiation into endothelial and mesodermal cell lineages was monitored using Flk-1 and Brachyury as the markers for respective cell lineages. (b) Semi-quantitative RT- PCR depicting the expressions of coding genes such as *Hoxc8* and *Hoxc6* as compared to that of *NC1, NC2* and *NC4*. The schematic shows the coding gene exons as black rectangles and intergenic transcripts as lines above the locus. Primer locations are denoted as pairs in dotted lines. (c) Semi- quantitative RT- PCR to compare expressions of *Hoxc8* and *NC4* in mES cells that are wild type or Mll null. (d) Quantitative RT-PCR to detect expressions of *Hoxc8* and *NC4* in wild type mES cells. Inset shows the low-level expressions of the two transcripts during first five days. (e) Both *Mll* and *Cyp33* are expressed throughout the course of differentiation.

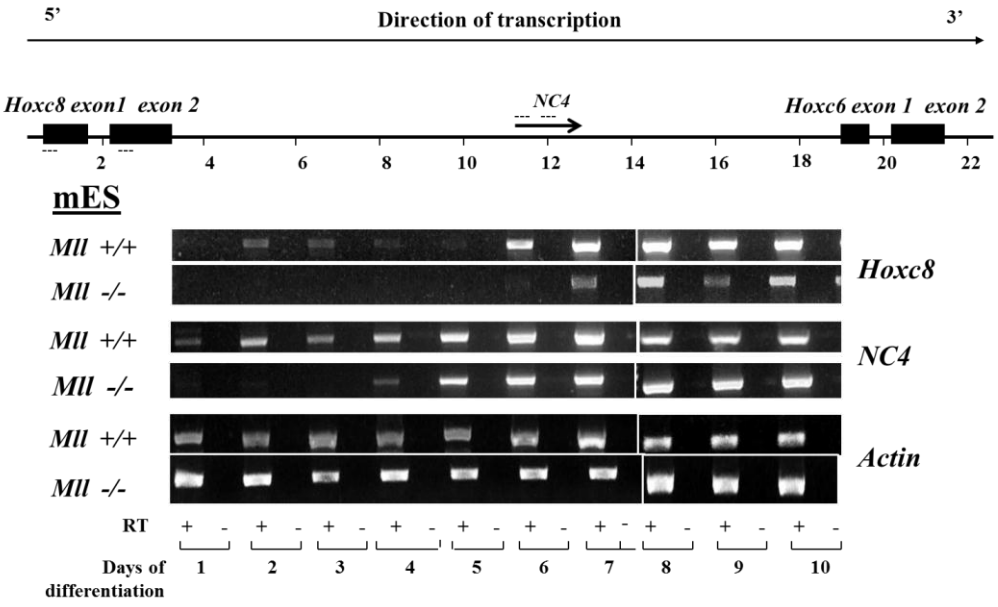
(a)


Undifferentiated
Day 2
Day 4
Day 6
Day 8
Day 10
Brachyury
Flk1
Actin
RT
+ - + - + - + - + - + - + - + - + - + -
Days of differentiation
1 2 3 4 5 6 7 8 9 10

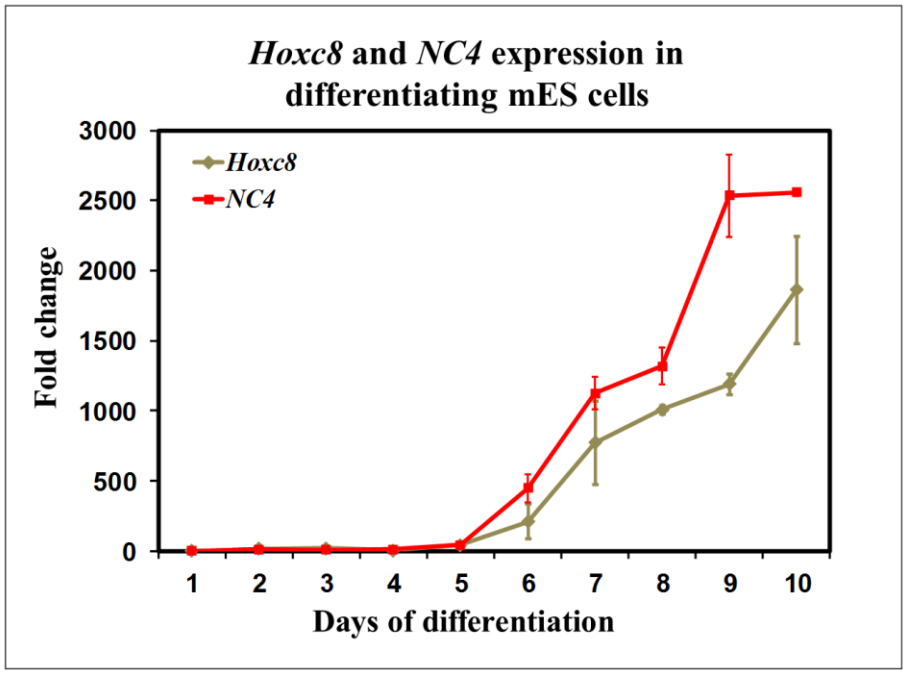
(b)


5'
Direction of transcription
3'
Hoxc8 exon1 exon 2
NC2
NC1
NC4
Hoxc6 exon 1 exon 2
2 4 6 8 10 12 14 16 18 20 22
NC4
NC2
NC1
Hoxc8
Hoxc6
Actin
RT
+ - + - + - + - + - + - + - + - + - + -
Days of differentiation
1 2 3 4 5 6 7 8 9 10

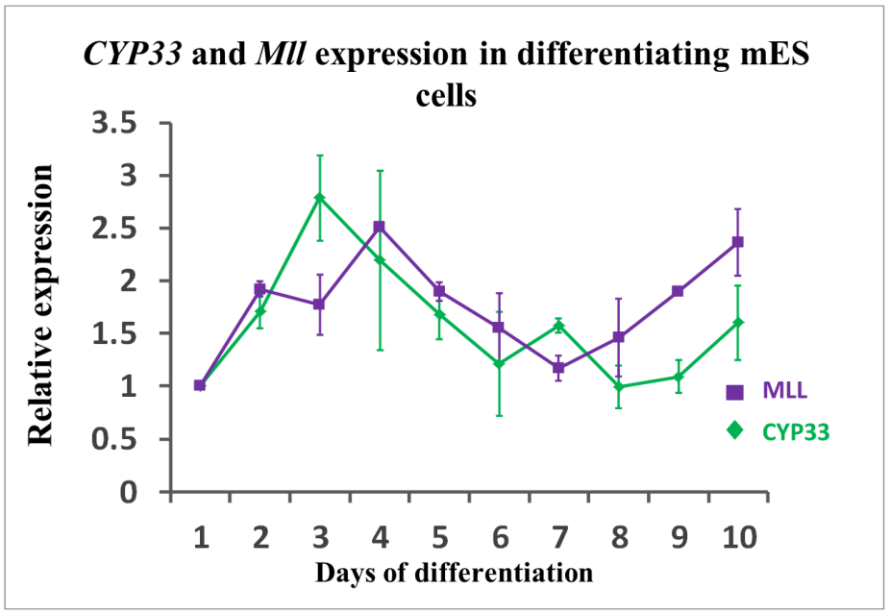
(c)


5'
Direction of transcription
3'
Hoxc8 exon1 exon 2
NC4
Hoxc6 exon 1 exon 2
2 4 6 8 10 12 14 16 18 20 22
mES
Mll +/+
Mll -/-
Hoxc8
NC4
Actin
RT
Days of differentiation
1 2 3 4 5 6 7 8 9 10

(d)


Hoxc8 and NC4 expression in differentiating mES cells
Hoxc8
NC4
Fold change
3000
2500
2000
1500
1000
500
0
1 2 3 4 5 6 7 8 9 10
Days of differentiation

(e)


CYP33 and *Mll* expression in differentiating mES cells
Relative expression
3.5
3
2.5
2
1.5
1
0.5
0
1
2
3
4
5
6
7
8
9
10
Days of differentiation
MLL
CYP33

In order to determine the effect of *Mll* knockout (Yu et al. 1995) on the expressions of *Hoxc8* and enhancer-overlapping *NC4*, a similar experiment was conducted in the *Mll* null mES cells. As shown in Fig. 12b, expressions of both *Hoxc8* and *NC4* in *Mll* null mES is delayed as compared to that in the wild type mES cells implying dependency of both the transcripts on *Mll*. It is not known whether Mll binds to the region governing *NC4* expression (e.g. putative promoter of *NC4*). The initial burst of *Hoxc8* transcription that is observed in the WT mES cells is lost in the *Mll* null mES cells and the expression corresponding to the second phase begins a day later. The overall levels of *Hoxc8* in the null cells are lower as compared to that in the WT cells. *NC4* expression is delayed by three days. The comparison of the expressions of *Hoxc8* and *NC4* in *Mll* null mES cells shows that *NC4* is still expressed earlier. The results suggest that Mll regulates both the timing and levels of expression of the two transcripts, and that the *NC4* transcript is expressed earlier than the *Hoxc8* transcript during EB development under Mll control or independent of it.

CYP33 binds a YAAUNY consensus RNA sequence motif.

The CYP33 RRM has only been partially characterized for its RNA binding properties (Mi et al. 1996, Wang et al. 2008, Hom et al. 2010, Wang et al. 2010, Park et al. 2010). None of the approaches undertaken represents an unbiased method of finding RNA ligands for CYP33. This study therefore addressed the RNA binding preference of CYP33 using an in vitro method of RNA selection from among a pool of randomized RNA sequences (see methods). SELEX (for Selective Evolution of Ligands by

Exponential Enrichment) is a fairly established method of finding high affinity oligonucleotide sequences binding to the protein of interest (Sakashita and Sakamoto 1994, Gopinath 2007). Recombinant GST tagged full length CYP33 or GST tagged CYP33 lacking the RRM (ΔRRM) [Fig. 13a] were immobilized onto glutathione-sepharose beads and used for binding to an initial pool of RNA (in the range of 10^{17} unique sequences), (Anderson et al. 2002; Djordjevic 2007) each containing a 30 nucleotide random sequence. The selected RNA sequences are then used in a subsequent round of PCR based amplification. The iterative steps in SELEX depend primarily on the affinity of the protein for a specific RNA sequence (Gopinath 2007). In this experiment, I considered the sequences bound by GST-ΔRRM CYP33 to be non-specific and used these in motif search and statistical analysis as background. After three rounds of selection by CYP33, 33 unique sequences were compared to identify a motif representing enriched RNA sequence. When compared, 33% of the total sequences selected contained a YAAUNY hexanucleotide motif [Fig. 13b] with the core AAU found at a significantly higher frequency ($p<0.025$) as compared to the background sequences selected by GST-ΔRRM CYP33 [Fig. 13c] as well as the frequency predicted from random nucleotide assortments. The rest of the 77% of the sequences when compared yielded a shorter version of the bound motif i.e. AAU at a significant frequency ($p< 0.025$) [Fig. 13b]. In order to determine if CYP33 has a preference for binding to a specific RNA secondary structure, the bound RNA sequences were studied for potential folding using the mfold algorithm written by the Zuker group (Zuker 2003) and compared for common secondary structure elements, distribution of YAAUNY and AAU in single versus double stranded

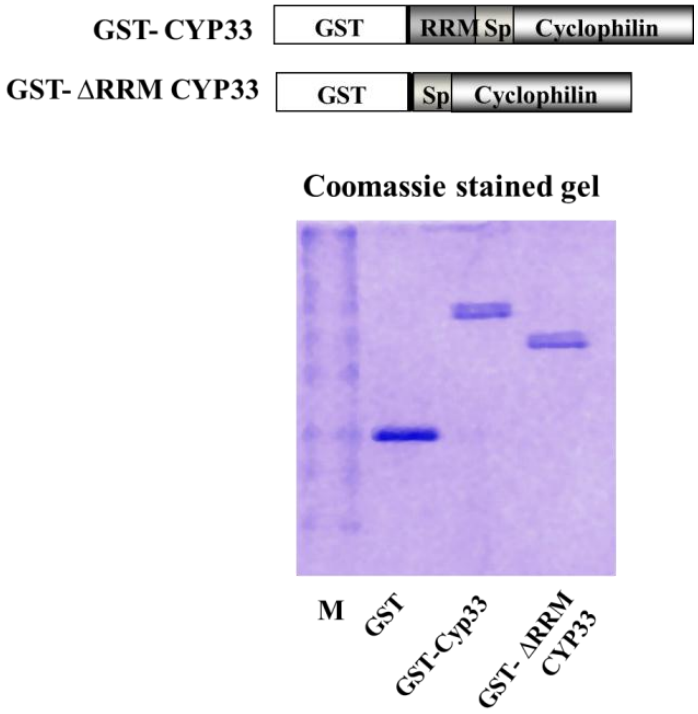
regions as well as occurrences within and outside of loops. Table 4 summarizes the results thus obtained. CYP33 prefers the YAAUNY motif in single stranded (ss) regions (72.39 % occurrence rate) whereas the AAU motif is found at almost equal rates in single as well as double stranded regions (52.18 %). Together, more than 70% of the AAU core motif in all the secondary structures compared was found to be single stranded suggesting that this may be the minimum core single stranded region necessary for binding to CYP33. Exemplifying the above are two stable folded RNA sequences found that contain the single stranded YAAUNY motif in loop and a ss core AAU in a bulge [Fig. 13d]. Most of the bound RNA folded to yield stable secondary structures with ΔGs ranging from -1.3 Kcal/mole through -12.3 Kcal/mole with an average of -5.19 Kcal/mole. In order to determine if the minimal AAU core consensus is required for binding to CYP33, RNA electrophoretic mobility shift assays were performed using GST tagged proteins and 5' end labeled RNA probes. As shown in [Fig. 14a], CYP33 binds an AAU containing oligonucleotide (see methods; Table 2) that folds into a single structure with ssAAU in the loop region. CYP33 also binds with high affinity to Poly A and Poly U, which was used as a positive control in this experiment. A control oligonucleotide selected from the ΔRRM CYP33 group that lacks AAU sequence does not bind or binds very weakly to CYP33. Neither GST alone nor any of the mutants of CYP33 (ΔRRM and L72P) bind RNA. A GST tagged MLLPHD3 was also tested for its binding to positive control Poly A RNA and was found not to bind Poly A. CYP33 binds both as a monomer and as a dimer to the tested RNA [Fig. 14a]. It is not very clear at this point whether this is due to the self-dimerization of CYP33 via its spacer region or due to an induced

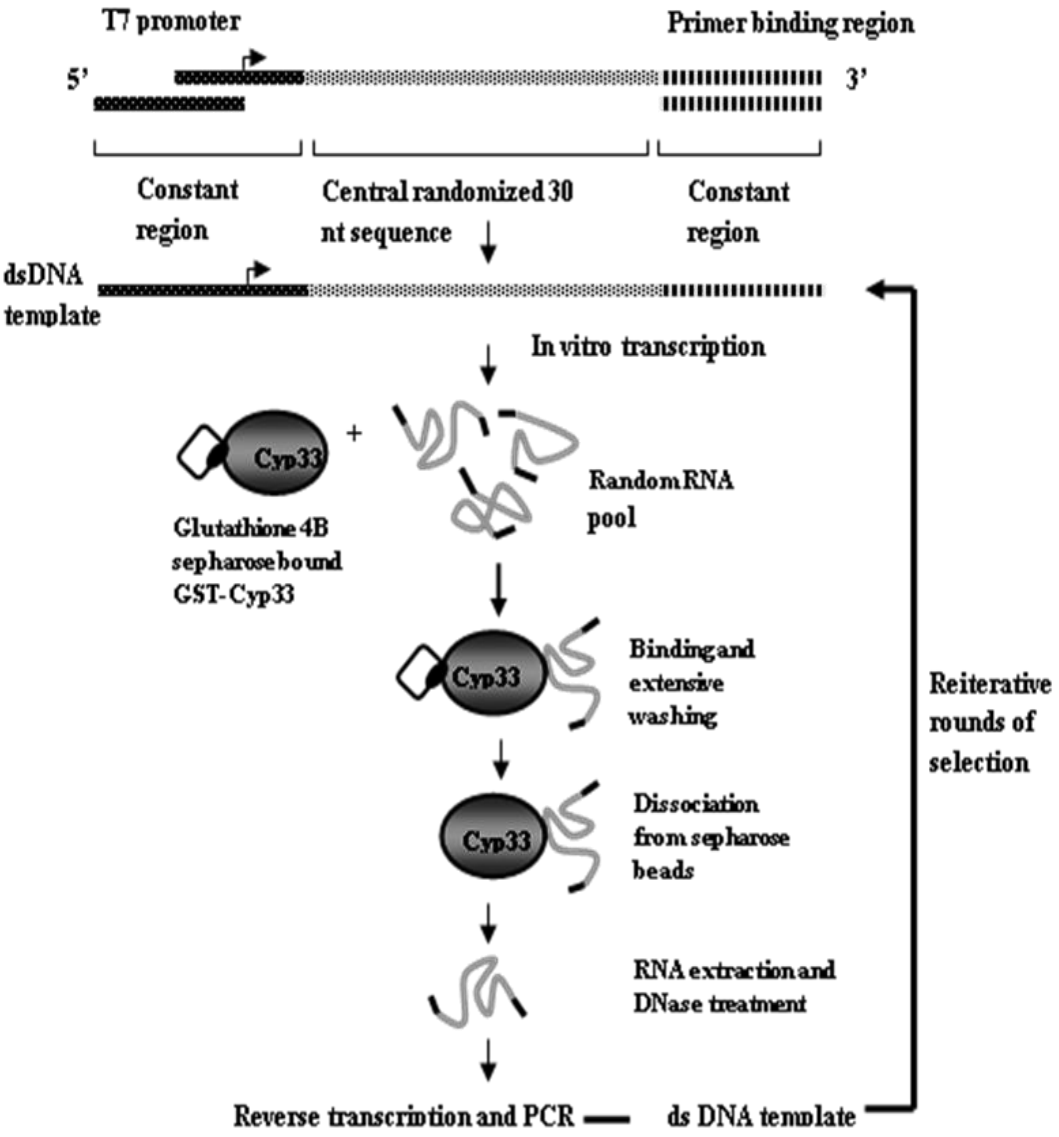
dimerization of CYP33 by binding to more than one site on an RNA sequence. Since the proportion of dimer is greater when bound to Poly A it is likely that the dimerization of RNA bound CYP33 is due to induction by a specific RNA sequence.

Figure 13: SELEX (Selective Evolution of Ligands by Exponential Enrichment) to find RNA ligand binding CYP33

(a) Constructs used for protein expression in SELEX experiment. Recombinant GST tagged proteins were expressed in bacteria and further purified. Coomassie staining was done for detection of protein and determination of purity. GST tagged ΔRRM CYP33 was used as a control whereas GST was used in the protocol at a step that preclears the prepared random RNA pool. The following schema shows the steps involved in SELEX (for details, see methods). (b) Three rounds of selection yielded a hexa nucleotide sequence of YAAUNY at a significant frequency as compared to a random sequence ($p < 0.025$) as well as to the control GST-ΔRRM CYP33 (shown in c). Found at a higher frequency was also a shorter version of AAU ($p < 0.025$). Sequences marked with an asterisk contain more than one binding motif. (c) Sequences selected by the control protein GST-ΔRRM CYP33 did not contain the YAAUNY at a significant frequency ($p = 0.55$). (d) Predicted secondary structure analysis using mFold shows a preferential binding of YAAUNY and AAU in single stranded region of folded RNA.

(a)


GST- CYP33
GST
RRM
Sp
Cyclophilin
GST- ΔRRM CYP33
GST
Sp
Cyclophilin
Coomassie stained gel
M
GST
GST-Cyp33
GST- ΔRRM CYP33


T7 promoter
Primer binding region
5'
3'
Constant region
Central randomized 30 nt sequence
Constant region
dsDNA template
In vitro transcription
Cyp33
+
Random RNA pool
Glutathione 4B sepharose bound GST-Cyp33
Cyp33
Binding and extensive washing
Cyp33
Dissociation from sepharose beads
RNA extraction and DNase treatment
Reverse transcription and PCR
ds DNA template
Reiterative rounds of selection

(b) Sequences bound by CYP33

Consensus:

(C) Sequences bound by ΔRRM CYP33 (conrol)

UAAUGCGUUAUCGGAACGUGUCCUGGCCAG
AUCAUUUAUAGCCCCUCAAAUACGGAUUGC
AUCAUUUAUAGCCCCUCAAAUACGGAUUGC
AGCGAUCUUCAAGAGUAUGCAUAAAUACAC
ACAUCAUUUAUAGCCCCUCAAAUACGGAUU
UAAACGUGCAAGAACUGUAAUAGACCGCUU
GAAAACCAAAAAUGAUCUCUGCCCCUUUCC
CCUGACUGCUCAAGGCUGAAACGCAUCAAC
UUACGAGACCCUCCUCUACACUUUAAAGAC
UAUUUUGCCCAACGUCCCACCCACGCAAGU
CCUGACUGCUCAAGGCUGAAACGCAUCAAC
ACUUACGAGACCCUCCUCUACACUUUAAAGAC
AAAGAUGUUCAAGAGUAUGCAUAANUACAC

(d)

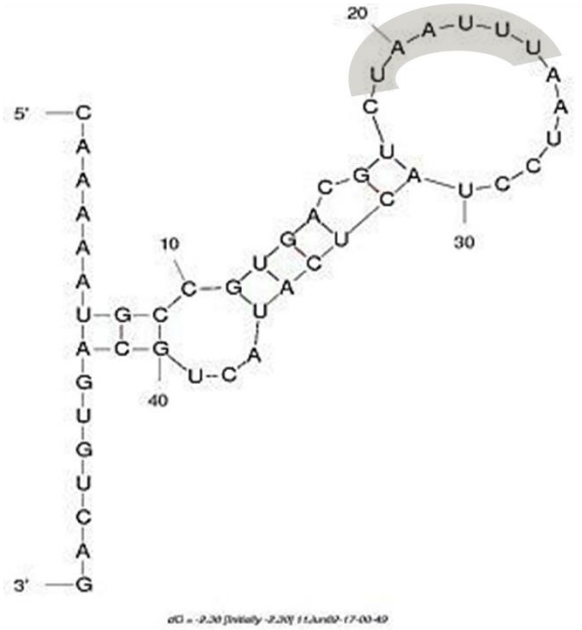
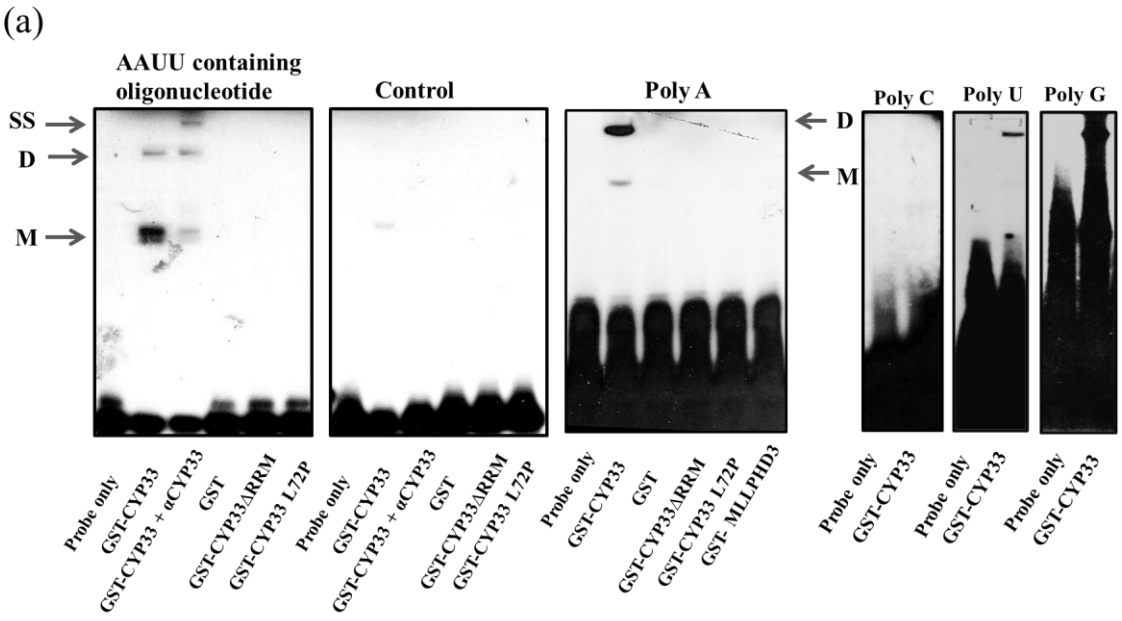
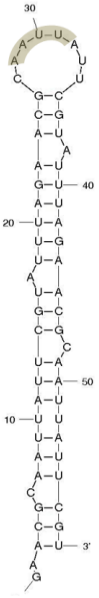


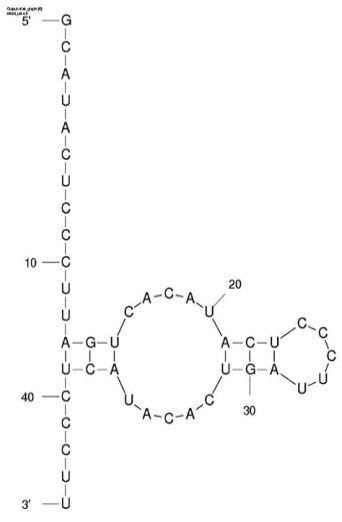
Figure 14: In vitro binding using RNA electrophoretic mobility shift assays. (a) This experiment was conducted using 5' end labeled 15 nucleotide long N5(AAUC)N6 oligonucleotide, control without the core AAU sequence from the GST-ΔRRM CYP33 group as well as Poly A RNA as a positive control. CYP33 binds AAUC containing oligonucleotide (AAUC oligo.) and Poly A. CYP33 binds AAUC oligo. to generate a monomer (M) and a dimer (D). CYP33 binds Poly A mostly as a dimer. Binding to the RNA is observed only with full length CYP33 and neither with a point mutant (L72P) that disrupts the β4 sheet nor with the ΔRRM. The MLL third PHD finger also does not bind Poly A indicating that the binding is specific to the RRM of CYP33. An antibody to CYP33 showed a supershifted band (SS) implying a specific binding between CYP33 and AAUC oligo. (b) Secondary structure of RNA (AAUC containing oligonucleotide and a control oligonucleotide) used in the RNA electrophoretic mobility shift assay.



b



AAUU containing oligo.



Control

Table 4: Secondary structure analysis of CYP33 enriched RNA sequences in a SELEX method. Between single and double stranded region, CYP33 prefers binding to single stranded motif. 52.18 % of CYP33 bound AAU motif is found in single stranded regions, whereas 72.39 % of YAAUNY motif is single stranded. Extending this analysis reveals the preference for binding either the long or short form of single stranded motif in the loop region.

	Whole or part of the motif, double stranded	Whole motif, single stranded within a loop	Whole motif, single stranded outside a loop
AAU	29.2 %	54.2 %	16.7 %
YAAUNY	35 %	39 %	27 %

Occurrence of the CYP33 binding sequence YAAUNY in the *HOXC8-HOXC6* intergenic region.

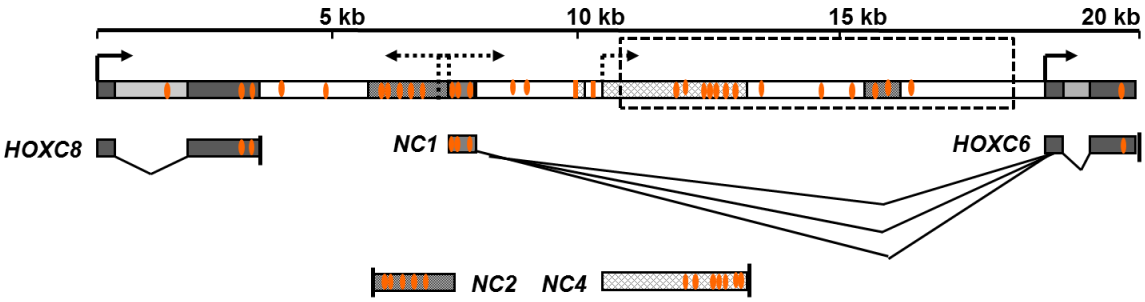
MLL dependent maintenance of *Hoxc8* expression depends on one or more regulatory elements located within a 8 kb long intergenic sequence (*Hoxc8 3'* RR) (Bradshaw et al. 1996).The *NC4* transcript overlaps this regulatory region of *Hoxc8* and has no significant protein coding potential. The *HOXC8-HOXC6* region, in both mouse and human, was scanned for the occurrence of YAAUNY motifs. As shown in Fig. 15a, the YAAUNY RNA motif is found in the *HOXC8* and *HOXC6* coding transcripts, albeit at a low frequency. The ncRNA such as the *NC2* and *NC4* on the other hand contain a higher density of YAAUNY with *NC2* featuring 6 of these motifs and *NC4* containing 7 of them. Supporting the results above is the observation of CYP33 binding to the endogenous *NC4* in the 293 cell line [Fig. 15b]. When compared for conservation among mammalian species using the PhastCons method of identifying conserved elements based on the phyloHMM (Hidden Markov Model) (Nielsen, Siepel, and Haussler 2005), *NC4* shows conservation of its 5' 1/3rd and 3' 1/3rd regions [Fig. 15a]. Incidentally, 5 out of the 7 YAAUNY motifs are present in the 3' conserved region of *NC4* suggesting its functional importance. In vivo binding of over expressed FLAG-CYP33 to endogenous protein coding transcripts such as *Hoxc8* and *Hoxc6* along with non-coding transcripts *NC1*, *NC2* and *NC4* was tested using anti-FLAG antibody in RNA immunoprecipitation without cross linking. As shown in Fig. 15b, CYP33 binds endogenous *NC2, NC3* and *NC4*, each containing 6, 2 and 7 YAAUNY motifs respectively. Binding is not seen with the endogenous *Hoxc8* transcript. These results imply that YAAUNY containing

endogenous transcripts can bind CYP33 in vivo despite the fact that these RNA represent low copy number cellular transcripts.

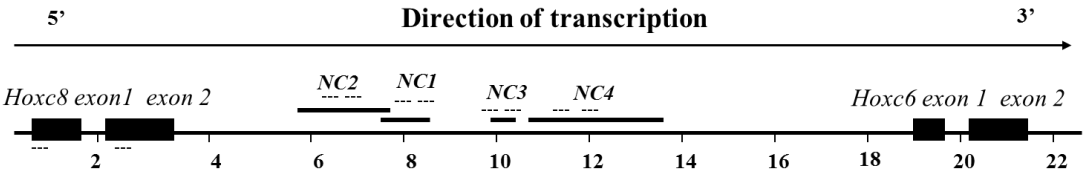
Figure 15: The YAAUNY motif is found at a greater density in the *NC4* transcript from the 3' regulatory region of *HOXC8*.

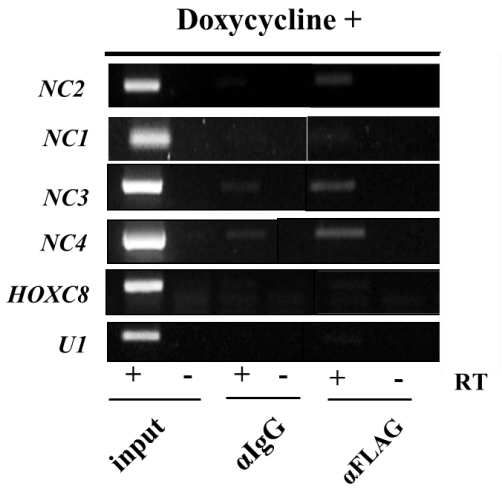
(a) *HOXC8* resides in the *HOXC* cluster on human chromosome 12 and mouse chromosome 15. The *Hoxc8* 3' RR described before is shown by the dotted rectangle. The *HOXC8-HOXC6* intergenic region is characterized by the presence of transcripts with little or no protein coding potential. These are denoted *NC1, NC2,* and *NC4* in the schema below. Since the *NC4* transcript overlaps with the 3' RR of *Hoxc8*, it may have a role in its regulation. The orange ovals represent the YAAUNY motif found throughout the intergenic region. The transcribed regions are denoted in grey whereas the non-transcribed regions are shown in white. Very few CYP33 binding motifs are found present in the intronic regions of different genes. *NC1* is found spliced with the exon 1 of *HOXC6* at three different locations as shown and contains YAAUNY motifs. The transcription start sites (TSS) for the protein coding genes is shown by solid arrows whereas putative TSS for ncRNA are denoted by dotted arrows. (b) In an inducible HEK 293 cell line overexpressing CYP33, endogenous ncRNA such as *NC1* and *NC4* are bound by the overexpressed CYP33 in a RNA immunoprecipitation assay.

(a)


5 kb
10 kb
15 kb
20 kb
HOXC8
NC1
HOXC6
NC2
NC4

(b)


5'
Direction of transcription
3'
Hoxc8 exon1
exon 2
NC2
NC1
NC3
NC4
Hoxc6 exon 1
exon 2
2
4
6
8
10
12
14
16
18
20
22


Doxycycline +
NC2
NC1
NC3
NC4
HOXC8
U1
+
-
+
-
+
-
RT
input
αIgG
αFLAG

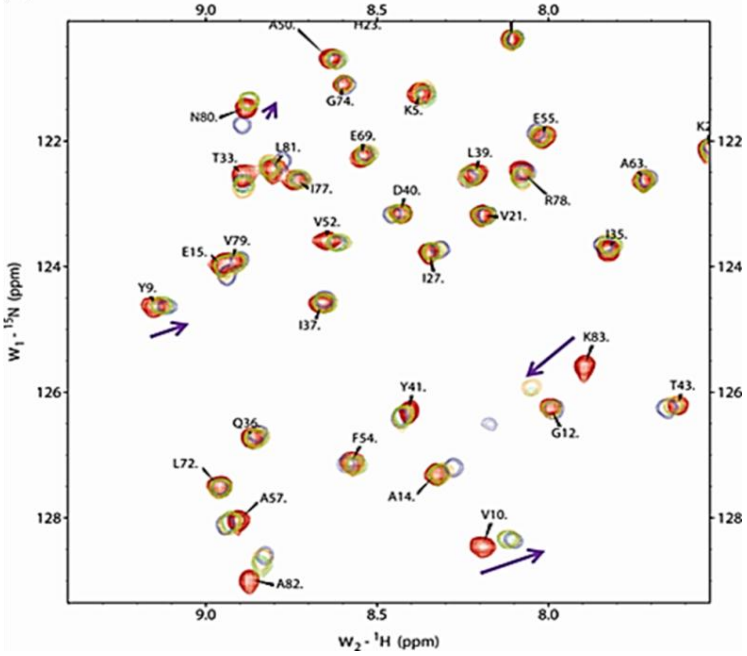
RNA binding and MLLPHD3 binding surfaces on the RRM of CYP33 overlap.

With the purpose of identifying RRM residues involved in an interaction with MLLPHD3, we collaborated with the laboratory of Dr. Bushweller from the University of Virginia who tested binding between Poly A (as a positive control), N5(AAUC)N6 RNA and a control without the core AAU sequence [Fig. 16a] and checked for perturbations in the backbone amide groups of amino acid residues from the RRM of CYP33 using NMR. Only select residues of the RRM showing shifts upon binding to various RNA are shown in the figure. As noticed, Y9 of RNP2 from CYP33 RRM shifts upon binding to PolyA as well as the N5(AAUC)N6 RNA oligonucleotide. In addition to Y9, both F49 and F51 from the RNP1 bind Poly A and AAUC containing oligonucleotide (data not shown). An overlap between RNA and MLLPHD3 interaction surfaces will be indicative of a competitive binding between CYP33 and its two interaction partners. Shown in Fig. 16b, are the interaction surfaces of CYP33 RRM with RNA and with MLLPHD3 as determined from the NMR data. Indeed there is an extensive overlap between the two interaction surfaces suggesting competitive binding by RNA and the MLLPHD3.

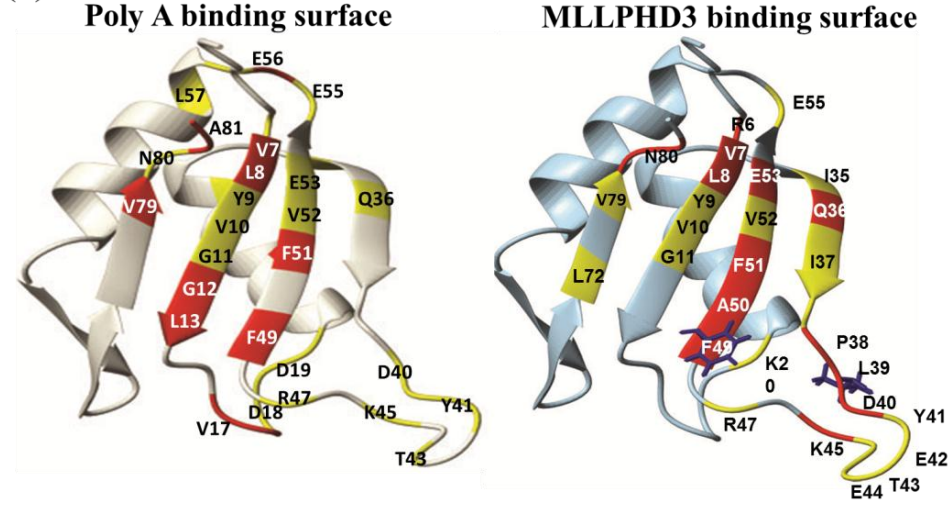
Figure 16: The RRM of CYP33 shares binding residues with RNA (PolyA and AAUC) and MLLPHD3.

(a) Selected regions of 15N-1H HSQCs of the Cyp33-RRM domain are shown. RRM alone (red), and in complex with N5(AAUC)N6 (green), Control oligo. N15 (orange) and Poly A (blue) (see also methods for sequences). (b) Shown below are the NMR shifts corresponding to amino acid residues of the RRM domain of CYP33 upon interaction with RNA and MLLPHD3. There is an extensive overlap between the interaction surfaces suggesting a competitive binding between the RNA (poly A ribonucleotide) and MLL. The residues in red represent greater shifts (> 0.2 ppm) whereas the yellow residues represent the weaker perturbations.

(a)


9.0
8.5
8.0
122
124
126
128
W1 - 15N (ppm)
W2 - 1H (ppm)
A50.
H23.
G74.
K5.
N80.
E55.
E69.
L81.
T33.
I77.
L39.
A63.
R78.
D40.
V21.
V52.
I35.
E15.
V79.
I27.
Y9.
I37.
K83.
Y41.
T43.
G12.
Q36.
F54.
L72.
A14.
A57.
V10.
A82.

(b)


Poly A binding surface
MLLPHD3 binding surface
E56
L57
E55
A81
V7
N80
L8
E53
Y9
Q36
V79
V52
V10
G11
F51
G12
L13
F49
D19
D40
R47
Y41
D18
K45
V17
T43
R6
E55
N80
V7
L8
E53
I35
V79
Y9
V52
Q36
V10
G11
F51
I37
L72
A50
F49
P38
K2
0
L39
D40
Y41
R47
K45
E42
E44
T43

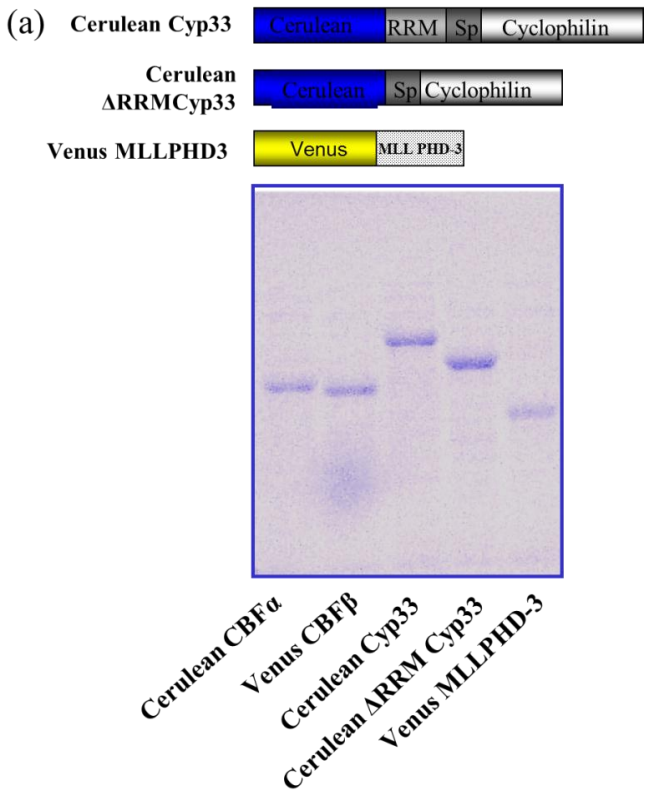
Poly A and YAAUNY containing RNA can competitively disrupt MLLPHD3 binding to CYP33.

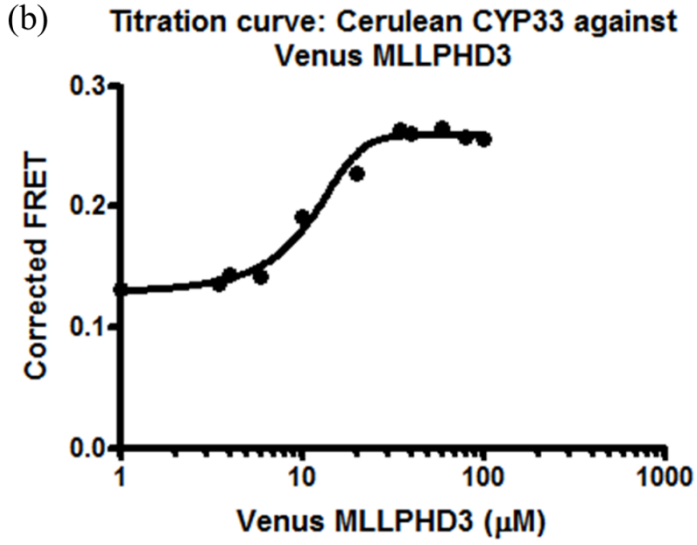
In the context of MLL mediated regulation of *Hoxc8*, CYP33 promotes the recruitment of co-repressors to MLL which correlates with MLL target gene repression (Xia et al. 2003). The switch from a repressed into an expressed state may be brought about by sequestration of CYP33 from the MLL complex by ncRNA thereby relieving repression at the locus. To test the above hypothesis, I carried out an in vitro experiment to test if YAAUNY containing RNA could competitively disrupt CYP33 binding to the MLLPHD3. Forster Resonance Energy Transfer (FRET) is a method that uses fluorophore (or fluorescent protein) tagged proteins to study an interaction between two molecules measuring the fluorescent energy transfer between the tagged proteins (Berney and Danuser 2003). As shown in Fig. 17a, either a full length CYP33 fused to Cerulean (donor cyan fluorescent protein) or its ΔRRM form fused to Cerulean was tested for corrected FRET or cFRET (see methods) when mixed with a MLLPHD3 peptide fused to Venus (acceptor yellow fluorescent protein). The proteins were used in a molar ratio of 4:35 (μM) based on a titration curve obtained for Cerulean-CYP33 and varying amounts of Venus-MLLPHD3 respectively [Fig. 17b] and also on an observation that the FRET efficiency increases with a decrease in the donor to acceptor ratio (Berney and Danuser 2003). The cFRET obtained for Cerulean-CYP33 and Venus tagged MLLPHD3 was significantly greater (by 64.75% with $p < 0.005$) than the background signal obtained with Cerulean-ΔRRM CYP33 and Venus- MLLPHD3 used as a control [Fig. 17c]. Different RNA sequences used in equimolar amount to MLLPHD3 were tested for the disruption

of the interaction between CYP33 and MLLPHD3 indicated by a decrease in cFRET. Only Poly A (used as a positive control for RNA binding by CYP33), a polynucleotide containing four YAAUNY motifs in a Poly C backbone $(CAAUNC)_4$, and the *NC4-4* fragment from the *NC4* ncRNA could decrease the cFRET obtained from the CYP33 and MLLPHD3 interaction to a significant extent (Poly A by 36.98%, $CAAUNC_4$ by 40.66% and NC4-4 by 35.78%). Even though both NC4-2 and NC4-3 fragments each carry one CYP33 binding hexanucletotide motif, no significant decrease in the cFRET from the interacting proteins was observed suggesting that one site may not be enough for observing a disruption of the interaction between CYP33 and MLLPHD3 using this assay. Poly C RNA was used as a negative control, and produced no reduction in the cFRET levels. The same set of RNA sequences were tested with the control pair Cerulean-ΔRRM CYP33 and Venus-MLLPHD3. Neither of the RNA sequences tested affects the background signal obtained from this control pair of proteins implying that the cFRET reduction observed with the full length CYP33 involved the RRM-specific interaction. MLLPHD3 has been demonstrated not to bind to Poly A RNA [Fig. 14a] further supporting the view that disruption of the CYP33-MLLPHD3 interaction occurs via RNA binding to the RRM. In order to ensure that the RNA mediated decrease in cFRET was not a non-specific effect, Poly A was tested for binding to non-RRM proteins CBFα and CBFβ that have been demonstrated earlier (Gorczynski et al. 2007) to be high affinity interaction protein partners. No decrease in cFRET was obtained when Poly A was incubated with Cerulean-CBFα and Venus-CBFβ [Fig. 17e].

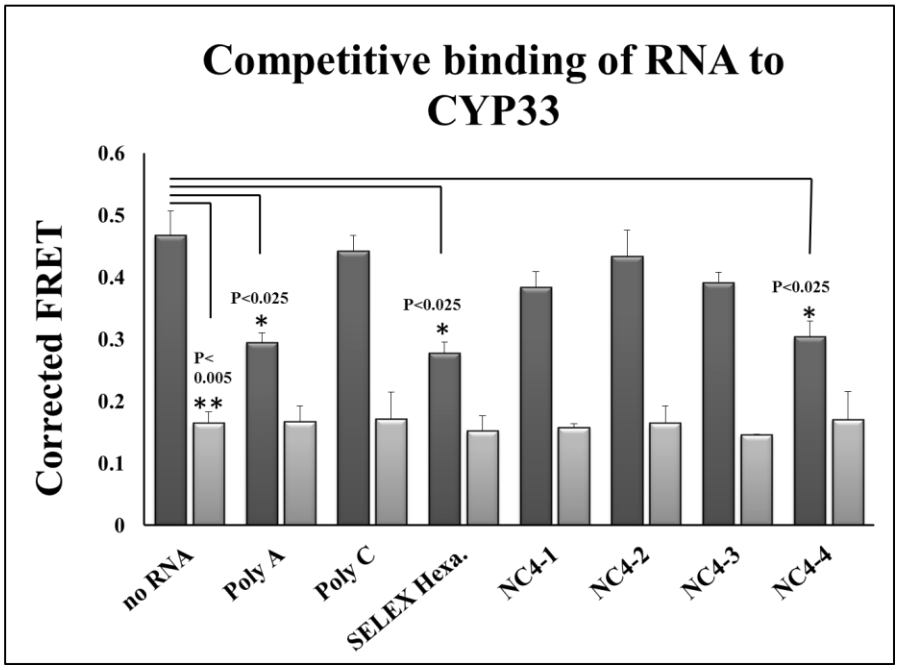
Figure 17: Poly A or YAAUNY containing RNA can competitively disrupt MLL-PHD3 binding to CYP33.

(a) Forster Resonance Energy Transfer was performed using either Cerulean tagged full length CYP33 or Cerulean tagged ΔRRM CYP33 with Venus tagged MLL-PHD3. Shown are the constructs made for the experiment and the proteins obtained from their expression in bacteria. (b) Titration was done using constant (4 μM) Cerulean-CYP33 and increasing amounts of Venus- MLLPHD3 (0-100 μM). The determined Kd for the binding between the two proteins was 10.97 ± 2.7μM. (c) Only Poly A (positive control), $CAAUCC_4$ or the NC4-4 RNA (*NC4* 3' region- see Fig. 16e also) could disrupt the corrected FRET signal generated from an interaction between Cerulean-CYP33 and Venus-MLLPHD3. No disruptive effect was seen with any of the RNA sequences on the Cerulean-ΔRRM-CYP33 and Venus-MLLPHD3 pair. (d) *NC4* overlaps with the *Hoxc8* 3'RR and contains 7 CYP33 binding sites. For the FRET based competitive assays, *NC4* was divided into four approximately equal length (approx. 700 nts. each) fragments transcribed in vitro from the T7 polymerase promoter. As noted, the 3' most fragment of *NC4* contains the highest density of YAAUNY motifs. (e) Corrected FRET between two heterodimerizing partners CBFα and CBFβ in the absence and presence of PolyA RNA.

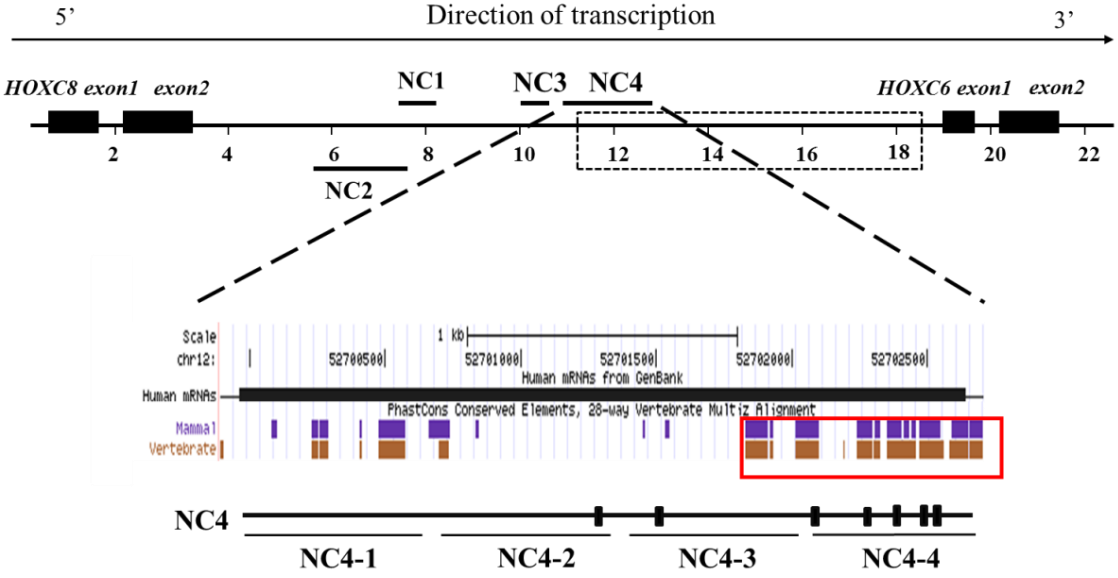

(a)
Cerulean Cyp33
Cerulean
RRM
Sp
Cyclophilin
Cerulean ΔRRMCyp33
Cerulean
Sp
Cyclophilin
Venus MLLPHD3
Venus
MLL PHD-3
Cerulean CBFα
Venus CBFβ
Cerulean Cyp33
Cerulean ΔRRM Cyp33
Venus MLLPHD-3


(b)
Titration curve: Cerulean CYP33 against Venus MLLPHD3
Corrected FRET
0.3
0.2
0.1
0.0
1
10
100
1000
Venus MLLPHD3 (μM)

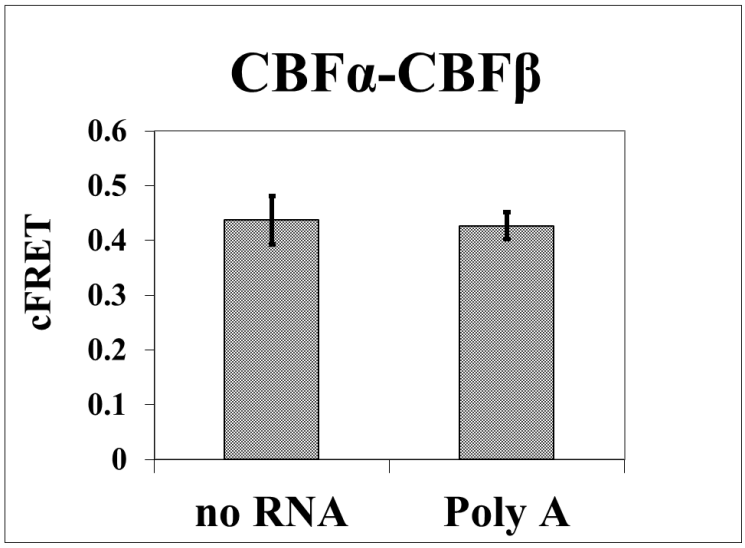
(c)


Competitive binding of RNA to CYP33
Corrected FRET
0.6
0.5
0.4
0.3
0.2
0.1
0
P< 0.005
**
P<0.025
*
P<0.025
*
P<0.025
*
no RNA
Poly A
Poly C
SELEX Hexa.
NC4-1
NC4-2
NC4-3
NC4-4

(d)


5'
Direction of transcription
3'
HOXC8 exon1 exon2
NC1
NC3 NC4
HOXC6 exon1 exon2
2 4 6 8 10 12 14 16 18 20 22
NC2
Scale
chr12:
1 Kb
52700500
52701000
52701500
52702000
52702500
Human mRNAs from GenBank
Human mRNAs
PhastCons Conserved Elements, 28-way Vertebrate Multiz Alignment
Mammal
Vertebrate
NC4
NC4-1
NC4-2
NC4-3
NC4-4

(e)



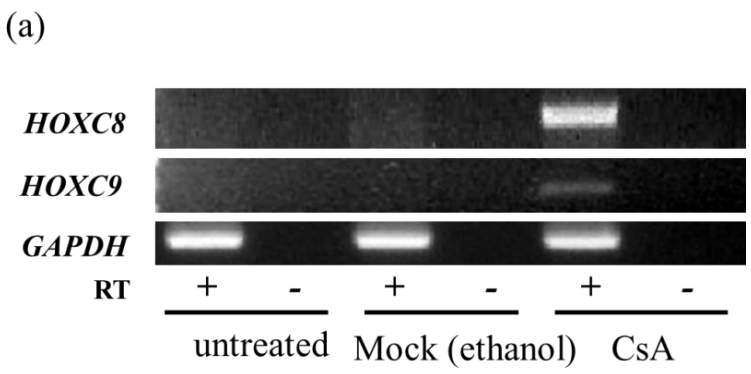
Expression of *NC4* in the MSA cell line induces expression of a previously silent *HOXC8* gene.

In order to test a possible role for *NC4* in *HOXC8* regulation, I tested if its ectopic expression can induce transcription of a previously silent *HOXC8* gene in a suitable cell line. MSA is an adherent cell line derived from a patient with human thyroid carcinoma and its entire *HOXC* locus is silent (Takahashi et al. 2004). In order to test if the MSA cell line expresses any of the *HOXC8-HOXC6* intergenic transcripts, quantitative RT-PCR was performed on total RNA from MSA. Neither *NC2* nor *NC4* was found expressed in MSA [Fig. 18b]. As anticipated, *HOXC8* and *HOXC6* were found silent in this cell line. Using both semi-quantitative and quantitative approaches, *NC1* and *NC3* transcripts were not found at detectable levels. I therefore performed semi-quantitative analysis and found neither *NC1* nor *NC3* expressed in these cells (data not shown). *MLL* and *CYP33* mRNAs, however, were found at detectable levels in a quantitative RT-PCR assay [Fig. 18b]. Since repression of the *HOXC* cluster in MSA cells may be due to higher levels of silencing mechanisms, I first tested if pharmacological inhibition of the endogenous CYP33 prolyl-isomerase activity in MSA could induce expression of *HOXC8*. It has been demonstrated earlier that the cyclophilin domain of CYP33 is essential for its modulation of MLL function towards gene repression (Fair et al. 2001). I therefore treated the MSA cells with cyclosporin A (CsA) for 2 days and tested for expression of the MLL target genes *HOXC8* and *HOXC9*. Compared to controls, CsA could induce expression of *HOXC8* and *HOXC9* in MSA cells [Fig. 18b] implying that these cells can serve as a good model to test for transinduction of *HOXC8* with ectopic

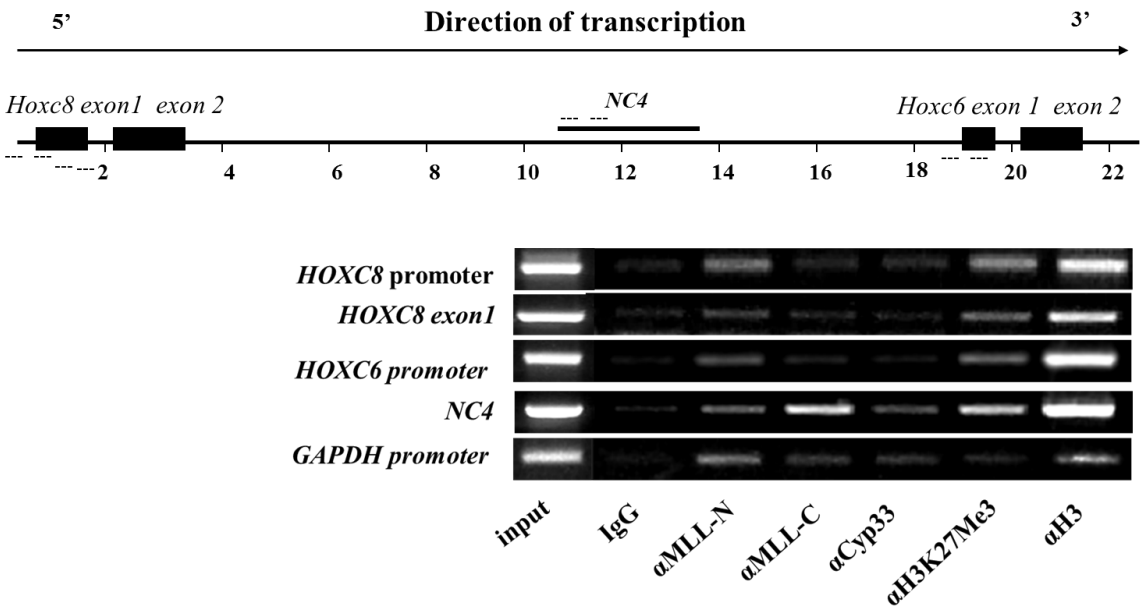
expression of intergenic transcripts. I further wanted to determine if MLL and CYP33 were bound to the *HOXC8* promoter in its silent state. Binding of MLL to a silent locus has not been reported before. ChIP assay was performed using MSA cells to determine if MLL-N, MLL-C and CYP33 were bound to the *HOXC8* promoter. Presence of H3K27me3 at the promoter was used as a positive control for silent loci. I found both MLL-N and MLL-C binding to the *HOXC8* promoter in the presence of H3K27me3 modification. The binding of MLL-C and CYP33 was found to be very weak. I have proved before that the expression of *NC4* in mouse MEFs is dependent on MLL [Fig.11a]. Binding of MLL-N, MLL-C and CYP33, therefore, was tested also at the putative promoter of *NC4*. I found a strong binding of MLL-C and CYP33 to *NC4* upstream region that may represent its putative promoter. The binding of MLL-N was weaker in the presence of H3K27me3 modification.

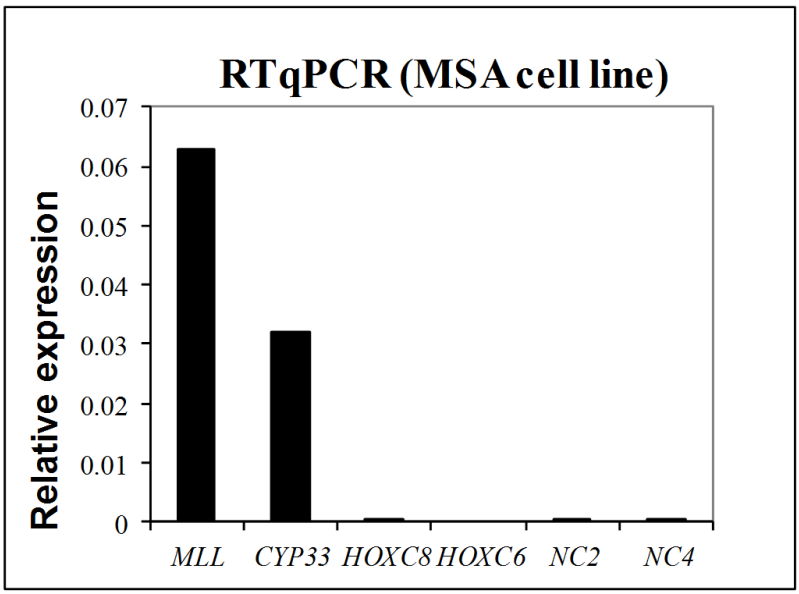
Figure 18: Ectopic expression of *NC4* in the MSA cell line induces expression of *HOXC8* in trans.

(a) MSA cells were treated with 1μg/ml of CsA for two days before testing for *HOXC8* and *HOXC9* expressions. Mock treatment involved treatment with ethanol (see methods) which is a vector used for CsA. (b) ChIP was performed on MSA cells to look for binding of MLL-N, MLL-C and CYP33 as well as H3K27me3 modification along with histone H3 at various loci mentioned. Primer locations to tested regions are denoted as dashed lines. The following panel denotes quantitative RT-PCR done to test for expressions of endogenous *MLL, CYP33, HOXC8, HOXC6, NC2* and *NC4*. (c) MSA cells were transiently transfected with pCMV-HA vector carrying cDNA for *NC3, NC4* (both with and without promoter as indicated). The expression of *HOXC8* in response to various plasmids was tested using quantitative RT-PCR.

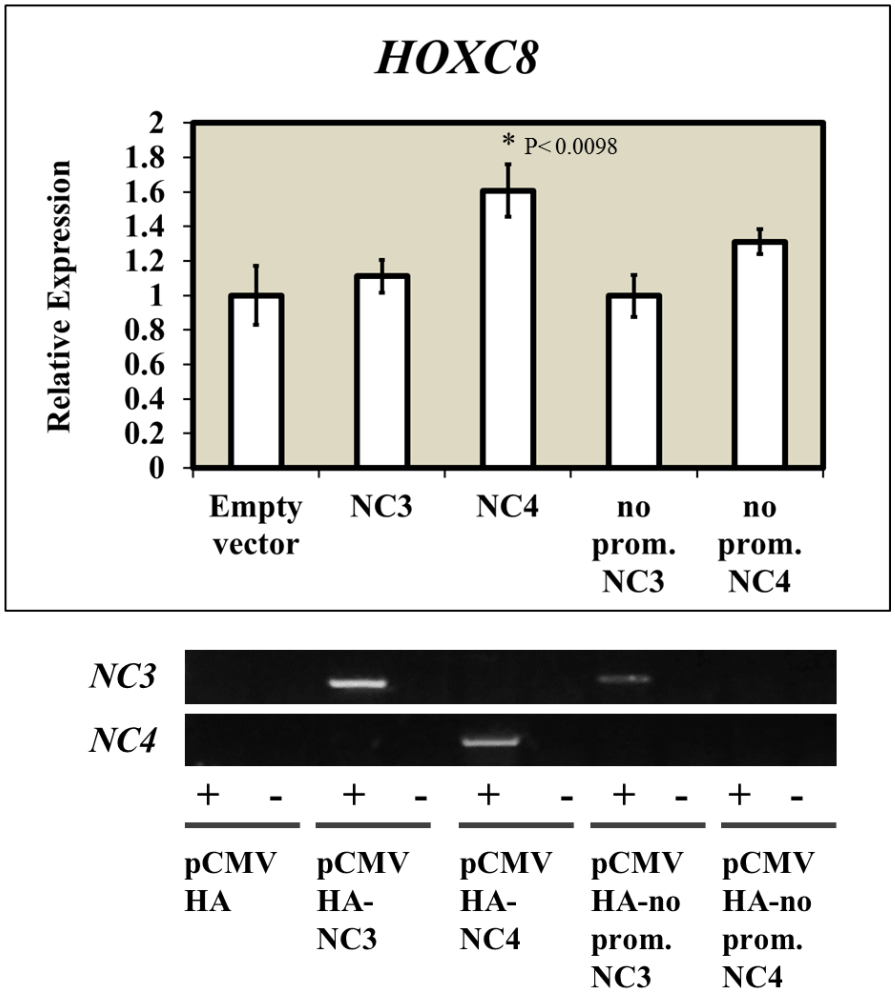


(b)


5'
Direction of transcription
3'
Hoxc8 exon1 exon 2
NC4
Hoxc6 exon 1 exon 2
2
4
6
8
10
12
14
16
18
20
22
HOXC8 promoter
HOXC8 exon1
HOXC6 promoter
NC4
GAPDH promoter
input
IgG
αMLL-N
αMLL-C
αCyp33
αH3K27Me3
αH3


RTqPCR (MSA cell line)
Relative expression
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
MLL
CYP33
HOXC8
HOXC6
NC2
NC4

(c)


HOXC8
Relative Expression
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
* P< 0.0098
Empty vector
NC3
NC4
no prom. NC3
no prom. NC4
NC3
NC4
+ - + - + - + - + -
pCMV HA
pCMV HA-NC3
pCMV HA-NC4
pCMV HA-no prom. NC3
pCMV HA-no prom. NC4

In support of this observation, I find a greater density of H3 at the silent genes such as *HOXC8, HOXC6* and *NC4* as compared to *GAPDH*, a housekeeping gene. When compared, there is a stronger binding of both MLL and CYP33 to *NC4* than to *HOXC8*. Having established that MSA cells lack expression of *HOXC8* and the intergenic transcripts *NC1*, *NC2*, *NC3* and *NC4* I ectopically expressed *NC3* and *NC4* in MSA cells to test if any of the transcripts could transinduce expression of *HOXC8* in the MSA cell line. *NC3* was included in this experiment even though it does not overlap with *Hoxc8* 3'RR mainly because it is mapped just downstream of the bivalent chromatin domain found in the mouse embryonic stem cells. As shown in [Fig. 18c], expression of *NC4* can induce expression of *HOXC8* (as compared to an empty vector control) by 60%. In order to distinguish between the possibilities of this effect being due to the *NC4* DNA sequence functioning as an enhancer that recruits co-activators to the *HOXC8* promoter or due its transcript, the same vector carrying *NC4* cDNA without a promoter was transfected into MSA cells. The *HOXC8* levels found in response to a promoter-less *NC4* plasmid were not found significantly greater than that with empty vector supporting the view that it is the *NC4* transcript that induces the expression of *HOXC8*. Furthermore, since this experiment involves a transient expression of *NC4*, the likelihood of plasmid integration into a genomic site causing *HOXC8* induction via a cis mechanism is less likely. The results above are consistent in the hypothesis that involvement of *NC4* in relieving repression at the *HOXC8* locus is mediated via sequestration of CYP33 by the *NC4* RNA.

CHAPTER FIVE

DISCUSSION

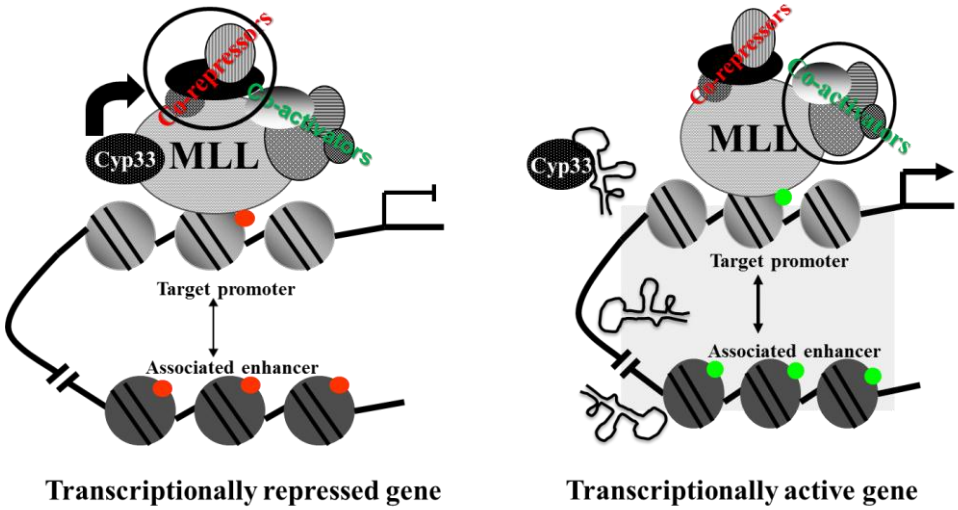
MLL or *Mixed Lineage Leukemia* gene participates in fusions with other genes after chromosome rearrangements, in human leukemia both de novo or after treatment with DNA topoisomerase II inhibitors used as cancer therapy for other malignancies (Harper and Aplan 2008). The MLL-fusion protein is leukemogenic and most patients afflicted with MLL leukemia have very poor prognosis (Marschalek 2011). Genetic deletion of MLL in mouse is embryonic lethal by day 10 of gestation (Yu et al. 1998; Hanson et al. 1999). The wild type MLL protein is a large multi-domain polypeptide capable of binding to both co-repressor and co-activator proteins (Xia et al. 2003; Ernst et al. 2001). It functions as a transcription maintenance regulator for numerous downstream genes including the type-I *Homeobox* (*Hox*) genes known to be involved in both embryogenesis as well as hematopoiesis (Deschamps et al. 1999). MLL is known to be required for the maintenance of *Hoxa* and *Hoxb* cluster genes during murine definitive hematopoiesis (Ernst et al. 2004; Argiropoulos and Humphries 2007). Of its many targets, *Hoxc8*, from the mammalian *Hoxc* cluster, has been the most characterized in terms of its expression

patterns and requirement during mouse embryogenesis (Belting, Shashikant, and Ruddle 1998). *Hox* gene expression both during mammalian as well as *Drosophila* embryogenesis occurs in two distinct phases identified based on their time and tissue specific expressions respectively (Belting, Shashikant, and Ruddle 1998, Deschamps et al. 1999). *Hoxc8* expression is also characterized by both an initiation phase as well as a maintenance phase each under the regulation of cis-regulatory elements present at the 5' end as well as 3' end of the *Hoxc8* gene respectively (Juan et al., 2003, Bradshaw et al., 1996). More importantly, the *Hoxc8* 3' regulatory region (3'RR) is required for the stabilization of *Hoxc8* expression specifically in the neural tube between the 5^{th} and 9^{th} somites as well as in mesoderm between the 9^{th} and 13^{th} somites (Juan et al., 2003, Bradshaw et al., 1996) after its expression has begun at the posterior end of the embryo thereby signifying its role in the maintenance of the *Hoxc8* expression in the mentioned tissues. In addition to the role of *Hoxc8* 3'RR, the expression of *Hoxc8* during mouse embryogenesis requires MLL function also (Hanson et al. 1999, Yu et al., 1998). Deletion of MLL early during embryogenesis causes loss of *Hox* gene expression after the initiation phase suggesting a functional relevance of the role of MLL and the cis-regulatory elements identified for *Hoxc8* (Yu et al., 1998).

The current study addresses molecular mechanism of MLL mediated regulation of the late phase expression of its target gene *Hoxc8* and the function of the *Hoxc8* 3'RR regulatory region (described earlier in Fig. 5). The *Hoxc8-Hoxc6* intergenic region that houses the *Hoxc8* 3'RR enhancer is transcribed into non-coding intergenic transcripts [Fig. 5]. Particularly, the *NC4* ncRNA arises from the *Hoxc8* 3'RR and its expression

positively correlates with that of the *Hoxc8* in both human and mouse cell lines [Fig. 9 b and c], mouse embryos [Fig. 9 d] as well as in the in vitro differentiating mouse embryoid bodies [Fig. 12 b and c]. MLL exists in a complex with other chromatin regulating proteins including different co-activators and co-repressors (Table 1). The interaction of CYP33 with MLL potentiates the co-repressor function at the MLL bound *HOXC8* locus causing its down regulation (Xia et al. 2003; Fair et al. 2001). This repression is mediated in part by H3 and H4 deacetylation (Dissertation studies by Mark Koonce). CYP33 interacts with both MLLPHD3 (Park et al. 2010; Fair et al. 2001, Hom et al. 2010; Wang et al. 2010) and RNA (Mi et al. 1996, Wang et al. 2008) via its RRM. We hypothesize that the *NC4* non-coding RNA transcribed from the *Hoxc8* 3'RR enhancer sequesters CYP33 from the MLL complex bound at the *Hoxc8* locus relieving repression and allowing for transcription from the *Hoxc8* promoter [Fig. 19].

Figure 19: Working model proposing a role of non-coding RNA in alleviation of CYP33 mediated repression at the *HOXC8* promoter through the disruption of the CYP33-MLL interaction. CYP33 interacts with MLLPHD3 to enhance binding of co-repressor proteins such as HDAC1 to the repression domain of MLL (left panel) which leads to a functional dominance of co-repressor proteins at the complex resulting in gene silencing. The *NC4* transcript contains CYP33 binding sites via which it competitively binds to CYP33 and sequesters it from the MLL complex. This in effect relieves the CYP33 mediated *HOXC8* repression - thereby allowing gene activation by co-activator proteins (right panel). The nucleosomes are depicted as light grey spheres at the *HOXC8* promoter whereas dark grey spheres represent nucleosomes at the enhancer region. The transcriptionally active promoter is shown by the arrow whereas the blunted line represents a silenced promoter. The stem and loop structures denote folded ncRNA *NC4* transcribed from the *Hoxc8* 3' RR. The small red and green spheres denote histone modifications for silenced and actively transcribing chromatin respectively.



This study identified a YAAUNY consensus RNA sequence motif [Fig. 13 b and d] that specifically binds CYP33 in vitro [Fig. 14 a] as well as in vivo [Fig. 15 b]. NMR studies from our collaborators Sangho Park and John Bushweller (University of Virginia) identified conserved aromatic amino acid residues Y9 and Y41 on the CYP33-RRM [Fig. 16 a] which are involved in an interaction with a CAAUCC RNA oligonucleotide. Some of the RNA interacting residues mapped on the CYP33 RRM overlap with those binding the MLLPHD3 [Fig. 16 b] implying that the binding by AAU and MLLPHD3 are mutually exclusive. Wang et al., (2008) describe preferential binding of a polyadenylation signal AAUAAA sequence to CYP33 (Wang et al. 2008) . Our method of identification of YAAUNY motif represents an unbiased approach of selection and enrichment of specific RNA sequence binding in vitro to CYP33. I find enrichment of the YAAUNY consensus motif in the 3' region of *NC4* [Fig. 15 a and Fig. 17 d] suggesting that it is a potential CYP33 binding sequence in endogenous transcripts. The *NC4* transcript sequence is highly conserved among mammals, especially in its 5' and 3' regions [Fig. 17 d]. Since *NC4* arises from the *Hoxc8* 3'RR enhancer, it is likely that the enhancer mediated regulation of *Hoxc8,* as proposed, involves a role of *NC4* in binding to CYP33 and titrating it away from the MLL complex. The proximal upstream region of *NC4* [Fig. 5] is characterized by the presence of a bivalent chromatin domain that contains both H3K4me3 and H3K27me3 in mouse embryonic stem cells (Bernstein et al. 2006). Bivalent chromatin has been identified in mammals in the context of developmental genes that are poised for transcription or silencing depending on their later cellular fate (Bernstein et al. 2006). Of the entire region encompassing *Hoxc8* and *Hoxc6*,

bivalent chromatin has been found to be present in the region between 7.6 kb to 10.6 kb relative to the *Hoxc8* transcription start site and not at the *Hoxc8* or *Hoxc6* promoters, clearly marking it as a developmentally important regulatory region (Suppl. Data; Bernstein et.al. 2006). Expression of the *NC4* transcript is found to precede and exceed that of *Hoxc8* during in vitro differentiation of mouse embryonic stem cells [Fig. 12 b and d] implying an early induction of the *NC4* region in response to differentiation cues. When comparing *MLL* wild type and *MLL* null mEBs, the expression of *NC4* was delayed by three days, whereas, the *Hoxc8* expression levels were both delayed by a day and decreased in *MLL* null cells indicating MLL dependency of both transcripts.

In order to critically assess if the binding of RNA and MLLPHD3 to CYP33 RRM is competitive, FRET was used with fluorescence protein tagged CYP33 and MLLPHD3. This method allowed us to prove that the steady state binding of Cerulean-CYP33 and Venus-MLLPHD3 is disrupted by YAAUNY containing RNA when added in equimolar amounts. As noted from Fig. 17c, positive control Poly A can disrupt the CYP33-PHD3 interaction while our negative control, polyC, cannot. Nevertheless an oligonucleotide with four copies of CAAUCC in a poly C backbone, or the NC4-4 RNA fragment containing 5 copies of the consensus YAAUNY sequence could each significantly disrupt the interaction between the MLLPHD3 and CYP33. Importantly this assay allowed for a direct testing of whether the NC4-4 fragment, which represents an endogenous nascent transcript sequence from humans, competitively binds to CYP33 in the presence of the MLLPHD3. The RNA mediated disruption of protein-protein interaction is specific to the CYP33-MLLPHD3 pair since no reduction in cFRET signal

is noted for the ΔCYP33-MLLPHD3 pair [Fig. 17e]. The current study also found binding of CYP33 to the endogenous *NC4* transcript despite its low abundance [Fig. 15b]. This result demonstrates a direct in vivo binding of CYP33 to endogenous non coding transcript *NC4*. A similar study pertaining to the dual recognition of MLLPHD3 and AAUAAA RNA sequence as binding partners by CYP33 RRM conducted by Hom et. al. (2010) also found that the binding surfaces for the two on the CYP33 RRM were overlapping (Hom et al. 2010). They further determined, using isothermal calorimetry, a dissociation constant of 1.9 μM for the CYP33-MLLPHD3 complex, and 198 μM for CYP33-AAUAAA RNA complex, suggesting a much higher affinity of CYP33 for MLLPHD3. Using NMR, they demonstrate that the chemical perturbations induced in the amino acids residues of CYP33 RRM upon binding of AAUAAA RNA are reversed upon addition of MLLPHD3. These results are not in agreement with our findings probably because the RNA sequences and the test conditions used by their group and this study are different. Furthermore, in our study we find NMR based shifts in the non-RNP residues from CYP33 RRM such as Y41, T43 and V52 [Fig. 16 a and b] that are located in the β2-β3 loop of CYP33 which were not identified by Hom et. al. when they studied the AAUAAA binding to the CYP33-RRM. The β2-β3 loop of CYP33 RRM is extended as compared to the same region from other known RRMs such as those of U2AF65 and may represent a CYP33-specific structure important for its molecular interactions. In our NMR assay we find Y41, T43 and V52 also shared for interaction with both MLLPHD3 and YAAUNY supporting the hypothesis of competitive binding of YAAUNY containing RNA, and the MLL-PHD3, to CYP33 [Fig. 17c]. Together these results

clearly demonstrate the potential role of the *NC4* transcript, arising from the 3' maintenance enhancer region of *Hoxc8,* in sequestration of CYP33 from the MLL complex. Similar to *Hoxc8,* transcription of *NC4* is found to be MLL dependent [Fig. 9c and Fig. 12] nevertheless it has not been elucidated if MLL and CYP33 bind to the upstream region of *NC4*. Using ChIP analysis in this study, both MLL and CYP33 were found to bind to the silent promoter of *HOXC8* in the MSA cell line [Fig. 18b]. The binding of CYP33 to the *HOXC8* promoter is very weak as compared to the control IgG and may be due to poor affinity of anti-CYP33 antibody to its epitope. The lack of *NC4* transcription in the MSA cell line correlates with the presence of the H3K27me3 histone modification as well as the binding of MLL and CYP33 to its upstream region, a putative *NC4* promoter [Fig 18b]. Enhancer mediated gene regulation in many cases involves long range interactions between the enhancer and promoter sequences with the intervening DNA looping out in both *Drosophila* (Ronshaugen and Levine 2004) and mammals (Ferraiuolo et al. 2010, Lee et al. 2010). In the current study I do not directly test for this possibility, however, I demonstrate that the *NC4* region contains the H3K27me3 histone modification, which has been recently identified as a part of signature of poised enhancers (Rada-Iglesias et al. 2011).

In order to test if the *HOXC* locus in MSA cells is irreversibly silenced, the MSA cells were first treated with Cyclosporin A (CsA) and tested for induction of *HOXC8* expression. As noticed in Fig. 18a, both the MLL targets *HOXC8* and *HOXC9* were transcriptionally reactivated upon CsA treatment suggesting that the repression on these genes is reversible. This reactivation may be related to inhibition of the CYP33 peptide-

prolyl isomerase activity by CsA. This result supports the previous observation in which an intact cyclophilin function has been demonstrated to be essential for the repression mediated by CYP33 in complex with MLL (Fair et. al., 2001). After verification of the reversible nature of *HOXC8* gene silencing in MSA cells, I ectopically expressed *NC4* transcript in these cells and tested for induction of the *HOXC8* gene. Upon ectopic expression of *NC4* from a transfected plasmid, [Fig. 18c] the endogenous *HOXC8* was induced up to 60% as compared to that of cells transfected with the empty vector control (average of three independent experiments) or with a plasmid containing a promoter-less *NC4* sequence. These results suggest that *NC4* functions as an enhancer-transcribed ncRNA that removes the repressive mechanisms from MLL bound *HOXC8* locus via sequestration of CYP33. A similar phenomenon of induction of intergenic transcription from β-globin locus in non-erythroid cells in response to genic transcription from plasmid-borne coding sequence has been demonstrated before, and termed 'transinduction' (Ashe et. al., 1997). This process does not require protein expression and the RNA transcribed from the transfected plasmid, in an in situ hybridization experiment, is observed to co-localize with the induced intergenic locus implying that the transcribed plasmid is brought in close proximity to the induced locus as has been proposed for enhancer-promoter interactions (Splintner 2006). Our system attempts testing for the involvement of a ncRNA transcribed from an enhancer element in the induction of transcription of a silent *HOXC8* gene. Even though *HOXC8* expression is observed to occur in response to ectopic expression of *NC4* [Fig. 18 c], it remains to be tested if the MLL and CYP33 bound endogenous *NC4* is itself transcribed in the transfected cells. An

alternative explanation to the induction of *HOXC8* in response to the ectopic expression of *NC4* in MSA cells could be an indirect mechanism that involves a direct binding of MLL to the ncRNA (Dinger et. al. 2008) and the recruitment of co-activator or transcription complexes to the *HOXC8* locus. I have demonstrated that MLLPHD3 by itself does not bind Poly A RNA [Fig. 14a] ruling out the possibility of MLL binding to RNA via this domain. MLL-N however, contains a less characterized snRNP domain (Hsieh et al. 2003) that may be involved in RNA binding thereby participating in a direct recruitment of MLL to the target gene locus or serving as a sensor of enhancer transcribed RNA and subsequent gene activation. It has also been proposed that the SET domain on MLL-C binds DNA and RNA (Krajewski et. al. 2005).

It is possible that the ncRNA mediated removal of repressive mechanisms at the target genetic loci is a more general phenomenon mediated by MLL-CYP33 containing complexes bound at other *HOX* genes or even non-*HOX* genes.

Non-coding RNAs have been increasingly found to form a newly identified layer of transcriptional regulation (Mattick 2009). In the context of human *HOX* genes, two long non-coding intergenic transcripts have been demonstrated to support both transcriptional activation and gene repression (Wang et al. 2011, Rinn et al. 2007). *HOTAIR*, transcribed from the *HOXC11-HOXC12* intergenic region complementary strand represses the entire *HOXD* locus in trans by directly recruiting the PRC2 complex protein SUZ12 to the target locus (Rinn et al. 2007). Another long intergenic non coding RNA (linc) RNA called HOTTIP transcribed from the complementary strand upstream of the 5’ most gene of the *HOXA* cluster functions in supporting transcription *HOXA* cluster

transcription by recruiting WDR5/MLL1 transcriptional activator protein complexes (Wang et al. 2011). These have been amongst the first few studies that attempted to determine the function of intergenic non-coding transcripts in *HOX* gene regulation. The *HOTAIR* RNA was further shown to directly interact also with co-repressor proteins such as EZH2 as well as LSD1/CoREST/REST complex thereby functioning as a scaffold for these proteins to mediate gene repression at the *HOXD* locus (Tsai et al. 2010). The ncRNA mediated regulation of chromatin at the target gene locus represents a more general process in which the mechanism behind the recruitment of ncRNA and co-repressor complexes to the target locus remains to be elucidated. Similar studies in *Drosophila* have led to identification of *bxd* ncRNA in the recruitment of trithorax group protein Ash1 to the *Ubx* promoter in tissues where this gene must be expressed (Sanchez-Elsner et al. 2006). A separate study; however, demonstrates using single cell resolution in situ hybridization as well as cell nuclei sorted for expression of a GFP-Ubx fusion protein, that the *bxd* transcription into ncRNA and *Ubx* transcription are mutually exclusive in the same cell and probably explained by transcriptional interference (Petruk et al. 2006). Our model proposes involvement of a gene specific enhancer transcribed ncRNA as a regulatory molecule that removes repressive proteins from the MLL target gene locus. Even though the current study has not tested for expression from *HOXC8* and *NC4* in single cells, the sense direction of transcription of the *NC4* transcription unit downstream of the *HOXC8* TSS [Fig. 9], precludes a direct transcriptional interference with *HOXC8* transcription.

Drosophila embryogenesis has provided ample evidences in the favor of co-binding of both Trithorax and Polycomb group of proteins to their binding elements (TRE/PRE) that are often found overlapping in the *Bithorax* complex (Maeda and Karch 2006; Ringrose and Paro 2007). During the progression of development, the function of one group of proteins dominates at a given locus thereby determining the gene expression state in the favor of expression or silencing (Ringrose and Paro 2007). The critical question concerning this mechanism is whether this process involves inhibition of function or removal of the opposing group of proteins. Studies done to map binding of the two groups of proteins on the *Drosophila* Bithorax complex in different cell lines with either active or silent *AbdB* gene using ChIP suggest that at least for some cell fates, the process of maintaining either of the gene expression states involves removal of proteins from the opposing group (Schwartz et al. 2010). Mammalian TRE/PRE have not been fully characterized yet. It is known that the repression domain of MLL can bind the Polycomb group of proteins such as the HPC2 and BMI1as well as the co-repressor protein HDAC1both in vitro and in vivo (Xia et al.). It yet remains to be determined if the Trithorax group protein MLL and the Polycomb group proteins HPC2 and BMI1 co-localize on the MLL target chromatin. Our model supports the view in which the switch of gene repression to transcription involves disruption of the binding of Polycomb group protein CYP33 (Andrew Dingwall; unpublished data) to the MLL complex. This mechanism; however, may be specific to regulatory pathway involving a role of transcribed enhancers in the modulation of transcription from their associated genes via interaction with their promoters using looping out of the intervening DNA. A similar

phenomenon has been visualized for iab5 enhancer and AbdB gene that demonstrates interaction between enhancer and promoter in the required embryonic segments (Ronshaugen and Levine 2004).

In summary, we have identified a consensus RNA sequence motif, YAAUNY, which preferentially binds CYP33. This sequence is found to be enriched in the *NC4* intergenic transcript, which arises from a previously identified *HOXC8* maintenance enhancer. During the course of development of in vitro differentiating mouse embryonic stem cells, we notice an early expression of *NC4* as compared to *Hoxc8* implying that the regulatory region encoding *NC4* undergoes activation of transcription early during development as compared to both *Hoxc8* and *Hoxc6*. We have provided evidence in favor of a competitive binding of specific RNA sequences containing the YAAUNY motif, with CYP33. We also show that the ectopic expression of the *NC4* RNA from a transfected plasmid results in reactivation of a previously silent *HOXC8* gene in a human cell line. These results are consistent with our central hypothesis that intergenic non-coding RNAs can regulate the transcriptional activity of MLL target genes by sequestering CYP33 from the promoter chromatin. This phenomenon of gene expression in response to an ectopic transcription of RNA from a transfected plasmid has been described earlier for the human β globin gene cluster, and called transinduction (Ashe et.al., 2001).

The transinduction evidence in the present study is very preliminary and will have to be tested more stringently with additional experiments to test if CYP33 is displaced from the promoter chromatin upon transinduction of *HOXC8* and whether it is bound to

the *NC4* nascent RNA during transinduction. The reduced binding of HDAC or other corepressors to MLL during transinduction by *NC4* also needs to be assessed. The preferential binding of CYP33 for specific RNA sequences should be further explored by using RNP immunoprecipitation to isolate natural RNAs that preferentially bind its RRM. In order to rule out the binding of chromatin regulating proteins such as EZH2, SUZ12, MLL etc. identified in other studies for binding to the regulatory ncRNA must be tested in our model for directly binding to *NC4* in vitro pull down of biotin tagged *NC4* followed by immunoblotting for the mentioned candidate proteins.

REFERENCES

Adler, H. T., R. Chinery, D. Y. Wu, S. J. Kussick, J. M. Payne, A. J. Fornace Jr, and D. C. Tkachuk. 1999. Leukemic HRX fusion proteins inhibit GADD34-induced apoptosis and associate with the GADD34 and hSNF5/INI1 proteins. *Molecular and Cellular Biology* 19 (10) (Oct): 7050-60.

Adler, Haskell T., Ferez S. Nallaseth, Gernot Walter, and Douglas C. Tkachuk. 1997. HRX leukemic fusion proteins form a heterocomplex with the leukemia-associated protein SET and protein phosphatase 2A. *Journal of Biological Chemistry* 272 (45) (11/07): 28407-14.

Akbari, O. S., A. Bousum, E. Bae, and R. A. Drewell. 2006. Unraveling cis-regulatory mechanisms at the abdominal-A and abdominal-B genes in the drosophila bithorax complex. *Developmental Biology* 293 (2) (May 15): 294-304.

Anand, S., W. C. Wang, D. R. Powell, S. A. Bolanowski, J. Zhang, C. Ledje, A. B. Pawashe, C. T. Amemiya, and C. S. Shashikant. 2003. Divergence of Hoxc8 early enhancer parallels diverged axial morphologies between mammals and fishes. *Proceedings of the National Academy of Sciences of the United States of America* 100 (26) (Dec 23): 15666-9.

Anderson, M., K. Fair, S. Amero, S. Nelson, P. J. Harte, and M. O. Diaz. 2002. A new family of cyclophilins with an RNA recognition motif that interact with members of the trx/MLL protein family in drosophila and human cells. *Development Genes and Evolution* 212 (3) (Apr): 107-13.

Ansari, K. I., S. Kasiri, I. Hussain, and S. S. Mandal. 2009. Mixed lineage leukemia histone methylases play critical roles in estrogen-mediated regulation of HOXC13. *The FEBS Journal* 276 (24) (Dec): 7400-11.

Argiropoulos, B., and R. K. Humphries. 2007. Hox genes in hematopoiesis and leukemogenesis. *Oncogene* 26 (47) (Oct 15): 6766-76.

Ashe, H. L., J. Monks, M. Wijgerde, P. Fraser and N J. Proudfoot, 1997. Intergenic transcription and transinduction of the human beta globin locus. Genes & Development (11) : 2494-1509.

Atchison, L., A. Ghias, F. Wilkinson, N. Bonini, and M. L. Atchison. 2003. Transcription factor YY1 functions as a PcG protein in vivo. *The EMBO Journal* 22 (6) (Mar 17): 1347-58.

Auweter, S. D., F. C. Oberstrass, and F. H. Allain. 2006. Sequence-specific binding of single-stranded RNA: Is there a code for recognition? *Nucleic Acids Research* 34 (17): 4943-59.

Ayton, P. M., and M. L. Cleary. 2003. Transformation of myeloid progenitors by MLL oncoproteins is dependent on Hoxa7 and Hoxa9. *Genes & Development* 17 (18): 2298-307.

Bae, E., V. C. Calhoun, M. Levine, E. B. Lewis, and R. A. Drewell. 2002. Characterization of the intergenic RNA profile at abdominal-A and abdominal-B in the drosophila bithorax complex. *Proceedings of the National Academy of Sciences of the United States of America* 99 (26) (Dec 24): 16847-52.

Bardine, N., C. Donow, B. Korte, A. J. Durston, W. Knochel, and S. A. Wacker. 2009. Two Hoxc6 transcripts are differentially expressed and regulate primary neurogenesis in xenopus laevis. *Developmental Dynamics : An Official Publication of the American Association of Anatomists* 238 (3) (Mar): 755-65.

Barges, S., J. Mihaly, M. Galloni, K. Hagstrom, M. Muller, G. Shanower, P. Schedl, H. Gyurkovics, and F. Karch. 2000. The fab-8 boundary defines the distal limit of the bithorax complex iab-7 domain and insulates iab-7 from initiation elements and a PRE in the adjacent iab-8 domain. *Development (Cambridge, England)* 127 (4) (Feb): 779-90.

Barna, M., T. Merghoub, J. A. Costoya, D. Ruggero, M. Branford, A. Bergia, B. Samori, and P. P. Pandolfi. 2002. Plzf mediates transcriptional repression of HoxD gene expression through chromatin remodeling. *Developmental Cell* 3 (4) (Oct): 499-510.

Belting, H. G., C. S. Shashikant, and F. H. Ruddle. 1998. Multiple phases of expression and regulation of mouse Hoxc8 during early embryogenesis. *The Journal of Experimental Zoology* 282 (1-2) (Sep-Oct 1): 196-222.

Berney, Claude, and Gaudenz Danuser. 2003. FRET or no FRET: A quantitative comparison. Abstract. *Biophysical journal* 84, no. 6:3992-4010.

Bernstein, B. E., T. S. Mikkelsen, X. Xie, M. Kamal, D. J. Huebert, J. Cuff, B. Fry, et al. 2006. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* 125 (2): 315-26.

Bieberich, C. J., M. F. Utset, A. Awgulewitsch, and F. H. Ruddle. 1990. Evidence for positive and negative regulation of the hox-3.1 gene. *Proceedings of the National Academy of Sciences of the United States of America* 87 (21) (Nov): 8462-6.

Bijl, J. J., J. W. van Oostveen, J. M. Walboomers, A. T. Brink, W. Vos, G. J. Ossenkoppele, and C. J. Meijer. 1998. Differentiation and cell-type-restricted expression of HOXC4, HOXC5 and HOXC6 in myeloid leukemias and normal myeloid cells. *Leukemia : Official Journal of the Leukemia Society of America, Leukemia Research Fund, U.K* 12 (11) (Nov): 1724-32.

Birney, E., S. Kumar, and A. R. Krainer. 1993. Analysis of the RNA-recognition motif and RS and RGG domains: Conservation in metazoan pre-mRNA splicing factors. *Nucleic Acids Research* 21 (25) (Dec 25): 5803-16.

Boncinelli, E., D. Acampora, M. Pannese, M. D'Esposito, R. Somma, G. Gaudino, A. Stornaiuolo, M. Cafiero, A. Faiella, and A. Simeone. 1989. Organization of human class I homeobox genes. *Genome / National Research Council Canada = Genome / Conseil National De Recherches Canada* 31 (2): 745-56.

Boyer, L. A., K. Plath, J. Zeitlinger, T. Brambrink, L. A. Medeiros, T. I. Lee, S. S. Levine, et al. 2006. Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature* 441 (7091) (May 18): 349-53.

Bradshaw, M. S., C. S. Shashikant, H. G. Belting, J. A. Bollekens, and F. H. Ruddle. 1996. A long-range regulatory element of Hoxc8 identified by using the pClasper vector. *Proceedings of the National Academy of Sciences of the United States of America* 93 (6) (Mar 19): 2426-30.

Brown, J. L., C. Fritsch, J. Mueller, and J. A. Kassis. 2003. The drosophila pho-like gene encodes a YY1-related DNA binding protein that is redundant with pleiohomeotic in homeotic gene silencing. *Development (Cambridge, England)* 130 (2) (Jan): 285-94.

Cai, H. N., D. N. Arnosti, and M. Levine. 1996. Long-range repression in the drosophila embryo. *Proceedings of the National Academy of Sciences of the United States of America* 93 (18) (Sep 3): 9309-14.

Carninci, P., T. Kasukawa, S. Katayama, J. Gough, M. C. Frith, N. Maeda, R. Oyama, et al. 2005. The transcriptional landscape of the mammalian genome. *Science (New York, N.Y.)* 309 (5740) (Sep 2): 1559-63.

Cernilogar, F. M., and V. Orlando. 2005. Epigenome programming by polycomb and trithorax proteins. *Biochemistry and Cell Biology = Biochimie Et Biologie Cellulaire* 83 (3) (Jun): 322-31.

Chen, J., D. A. Santillan, M. Koonce, W. Wei, R. Luo, M. J. Thirman, N. J. Zeleznik-Le, and M. O. Diaz. 2008. Loss of MLL PHD finger 3 is necessary for MLL-ENL-induced hematopoietic stem cell immortalization. *Cancer Research* 68 (15) (Aug 1): 6199-207.

Christophersen, N. S., and K. Helin. 2010. Epigenetic control of embryonic stem cell fate. *The Journal of Experimental Medicine* 207 (11) (Oct 25): 2287-95.

Cui, K., C. Zang, T. Y. Roh, D. E. Schones, R. W. Childs, W. Peng, and K. Zhao. 2009. Chromatin signatures in multipotent human hematopoietic stem cells indicate the fate of bivalent genes during differentiation. *Cell Stem Cell* 4 (1) (Jan 9): 80-93.

Cumberledge, S., A. Zaratzian, and S. Sakonju. 1990. Characterization of two RNAs transcribed from the cis-regulatory region of the abd-A domain within the drosophila bithorax complex. *Proceedings of the National Academy of Sciences of the United States of America* 87 (9) (May): 3259-63.

Daser, A., and T. H. Rabbitts. 2005. The versatile mixed lineage leukaemia gene MLL and its many associations in leukaemogenesis. *Seminars in Cancer Biology* 15 (3): 175-88.

Deschamps, J., E. van den Akker, S. Forlani, W. De Graaff, T. Oosterveen, B. Roelen, and J. Roelfsema. 1999. Initiation, establishment and maintenance of hox gene expression patterns in the mouse. *The International Journal of Developmental Biology* 43 (7): 635-50.

Dinger, M. E., P. P. Amaral, T. R. Mercer, K. C. Pang, S. J. Bruce, B. B. Gardiner, M. E. Askarian-Amiri, et al. 2008. Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation. *Genome Research* 18 (9) (Sep): 1433-45.

Dinger, M. E., K. C. Pang, T. R. Mercer, and J. S. Mattick. 2008a. Differentiating protein-coding and noncoding RNA: Challenges and ambiguities. *PLoS Computational Biology* 4 (11) (Nov): e1000176.

———. 2008b. Differentiating protein-coding and noncoding RNA: Challenges and ambiguities. *PLoS Computational Biology* 4 (11) (Nov): e1000176.

Djordjevic, M. 2007. SELEX experiments: New prospects, applications and data analysis in inferring regulatory pathways. *Biomolecular Engineering* 24 (2) (Jun): 179-89.

Dou, Y., T. A. Milne, A. J. Tackett, E. R. Smith, A. Fukuda, J. Wysocka, C. D. Allis, B. T. Chait, J. L. Hess, and R. G. Roeder. 2005. Physical association and coordinate function of the H3 K4 methyltransferase MLL1 and the H4 K16 acetyltransferase MOF. *Cell* 121 (6) (Jun 17): 873-85.

Duboule, D., and J. Deschamps. 2004. Colinearity loops out. *Developmental Cell* 6 (6) (Jun): 738-40.

Duboule, Denis. 1998. Vertebrate hox gene regulation: Clustering and/or colinearity? *Current Opinion in Genetics & Development* 8 (5) (10): 514-8.

Duncan, I. 1987. The bithorax complex. *Annual Review of Genetics* 21 (1) (12/01; 2011/08): 285-319.

EngstrÃ¶m, PÃ¤r,G., Harukazu Suzuki, Noriko Ninomiya, Altuna Akalin, Luca Sessa, Giovanni Lavorgna, Alessandro Brozzi, et al. 2006. Complex loci in human and mouse genomes. *PLoS Genet* 2 (4) (04/28): e47.

Erfurth, F. E., R. Popovic, J. Grembecka, T. Cierpicki, C. Theisler, Z. B. Xia, T. Stuart, M. O. Diaz, J. H. Bushweller, and N. J. Zeleznik-Le. 2008. MLL protects CpG clusters from methylation within the Hoxa9 gene, maintaining transcript expression. *Proceedings of the National Academy of Sciences of the United States of America* 105 (21) (May 27): 7517-22.

Ernst, P., J. K. Fisher, W. Avery, S. Wade, D. Foy, and S. J. Korsmeyer. 2004. Definitive hematopoiesis requires the mixed-lineage leukemia gene. *Developmental Cell* 6 (3) (Mar): 437-43.

Ernst, P., J. Wang, M. Huang, R. H. Goodman, and S. J. Korsmeyer. 2001. MLL and CREB bind cooperatively to the nuclear coactivator CREB-binding protein. *Molecular and Cellular Biology* 21 (7) (Apr): 2249-58.

Faber, J., A. V. Krivtsov, M. C. Stubbs, R. Wright, T. N. Davis, M. van den Heuvel-Eibrink, C. M. Zwaan, A. L. Kung, and S. A. Armstrong. 2009. HOXA9 is required for survival in human MLL-rearranged acute leukemias. *Blood* 113 (11) (Mar 12): 2375-85.

Fair, K., M. Anderson, E. Bulanova, H. Mi, M. Tropschug, and M. O. Diaz. 2001. Protein interactions of the MLL PHD fingers modulate MLL target gene regulation in human cells. *Molecular and Cellular Biology* 21 (10) (May): 3589-97.

Ferraiuolo, M. A., M. Rousseau, C. Miyamoto, S. Shenker, X. Q. Wang, M. Nadler, M. Blanchette, and J. Dostie. 2010. The three-dimensional architecture of hox cluster silencing. *Nucleic Acids Research* 38 (21) (Nov 1): 7472-84.

Gopinath, Subash. 2007. Methods developed for SELEX. *Analytical and Bioanalytical Chemistry* 387 (1) (-01-01/): 171,182; 182.

Gorczynski, M. J., J. Grembecka, Y. Zhou, Y. Kong, L. Roudaia, M. G. Douvas, M. Newman, et al. 2007. Allosteric inhibition of the protein-protein interaction between the leukemia-associated proteins Runx1 and CBFbeta. *Chemistry & Biology* 14 (10) (Oct): 1186-97.

Graveley, B. R. 2001. Alternative splicing: Increasing diversity in the proteomic world. *Trends in Genetics : TIG* 17 (2) (Feb): 100-7.

Gribnau, J., K. Diderich, S. Pruzina, R. Calzolari, and P. Fraser. 2000. Intergenic transcription and developmental remodeling of chromatin subdomains in the human beta-globin locus. *Molecular Cell* 5 (2) (Feb): 377-86.

Grimaud, C., N. Negre, and G. Cavalli. 2006. From genetics to epigenetics: The tale of polycomb group and trithorax group genes. *Chromosome Research : An International Journal on the Molecular, Supramolecular and Evolutionary Aspects of Chromosome Biology* 14 (4): 363-75.

Gruzdeva, N., O. Kyrchanova, A. Parshikov, A. Kullyev, and P. Georgiev. 2005. The mcp element from the bithorax complex contains an insulator that is capable of pairwise interactions and can facilitate enhancer-promoter communication. *Molecular and Cellular Biology* 25 (9) (May): 3682-9.

Gyurkovics, H., J. Gausz, J. Kummer, and F. Karch. 1990. A new homeotic mutation in the drosophila bithorax complex removes a boundary separating two domains of regulation. *The EMBO Journal* 9 (8) (Aug): 2579-85.

Haddad, F., A. X. Qin, P. W. Bodell, W. Jiang, J. M. Giger, and K. M. Baldwin. 2008. Intergenic transcription and developmental regulation of cardiac myosin heavy chain genes. *American Journal of Physiology.Heart and Circulatory Physiology* 294 (1) (Jan): H29-40.

Hagstrom, K., M. Muller, and P. Schedl. 1996. Fab-7 functions as a chromatin domain boundary to ensure proper segment specification by the drosophila bithorax complex. *Genes & Development* 10 (24) (Dec 15): 3202-15.

Hanson, R. D., J. L. Hess, B. D. Yu, P. Ernst, M. van Lohuizen, A. Berns, N. M. van der Lugt, et al. 1999. Mammalian trithorax and polycomb-group homologues are antagonistic regulators of homeotic development. *Proceedings of the National Academy of Sciences of the United States of America* 96 (25) (Dec 7): 14372-7.

Harper, D. P., and P. D. Aplan. 2008. Chromosomal rearrangements leading to MLL gene fusions: Clinical and biological aspects. *Cancer Research* 68 (24) (Dec 15): 10024-7.

Hayashizaki, Y., and P. Carninci. 2006. Genome network and FANTOM3: Assessing the complexity of the transcriptome. *PLoS Genetics* 2 (4) (Apr): e63.

Hess, J. L. 2004. MLL: A histone methyltransferase disrupted in leukemia. *Trends in Molecular Medicine* 10 (10): 500-7.

Ho, M. C., B. J. Schiller, S. E. Goetz, and R. A. Drewell. 2009. Non-genic transcription at the drosophila bithorax complex functional activity of the dark matter of the genome. *The International Journal of Developmental Biology* 53 (4): 459-68.

Hom, R. A., P. Y. Chang, S. Roy, C. A. Musselman, K. C. Glass, A. I. Selezneva, O. Gozani, R. F. Ismagilov, M. L. Cleary, and T. G. Kutateladze. 2010. Molecular mechanism of MLL PHD3 and RNA recognition by the Cyp33 RRM domain. *Journal of Molecular Biology* 400 (2) (Jul 9): 145-54.

Hou, Z., E. M. Kelly, and S. L. Robia. 2008. Phosphomimetic mutations increase phospholamban oligomerization and alter the structure of its regulatory complex. *The Journal of Biological Chemistry* 283 (43) (Oct 24): 28996-9003.

Hsieh, J. J., E. H. Cheng, and S. J. Korsmeyer. 2003. Taspase1: A threonine aspartase required for cleavage of MLL and proper HOX gene expression. *Cell* 115 (3): 293-303.

Jin, S., H. Zhao, Y. Yi, Y. Nakata, A. Kalota, and A. M. Gewirtz. 2010. c-myb binds MLL through menin in human leukemia cells and is an important driver of MLL-associated leukemogenesis. *The Journal of Clinical Investigation* 120 (2) (Feb 1): 593-606.

Juan, A. H., and F. H. Ruddle. 2003. Enhancer timing of hox gene expression: Deletion of the endogenous Hoxc8 early enhancer. *Development (Cambridge, England)* 130 (20) (Oct): 4823-34.

Kagey, M. H., J. J. Newman, S. Bilodeau, Y. Zhan, D. A. Orlando, N. L. van Berkum, C. C. Ebmeier, et al. 2010. Mediator and cohesin connect gene expression and chromatin architecture. *Nature* 467 (7314) (Sep 23): 430-5.

Kaufman, T. C., M. A. Seeger, and G. Olsen. 1990. Molecular and genetic organization of the antennapedia gene complex of drosophila melanogaster. *Advances in Genetics* 27 : 309-62.

Kim, J. D., A. K. Hinz, A. Bergmann, J. M. Huang, I. Ovcharenko, L. Stubbs, and J. Kim. 2006. Identification of clustered YY1 binding sites in imprinting control regions. *Genome Research* 16 (7) (Jul): 901-11.

Kong, L., Y. Zhang, Z. Q. Ye, X. Q. Liu, S. Q. Zhao, L. Wei, and G. Gao. 2007. CPC: Assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Research* 35 (Web Server issue) (Jul): W345-9.

Krajewski,W., T. Nakamura, A. Mazo, and Eli Canaani. 2005. *Mol. Cell. Biology* 25(5): 1891-1899

Kroon, E., J. Krosl, U. Thorsteinsdottir, S. Baban, A. M. Buchberg, and G. Sauvageau. 1998. Hoxa9 transforms primary bone marrow cells through specific collaboration with Meis1a but not Pbx1b. *The EMBO Journal* 17 (13) (Jul 1): 3714-25.

Kwon, Y., J. Shin, H. W. Park, and M. H. Kim. 2005. Dynamic expression pattern of Hoxc8 during mouse early embryogenesis. *The Anatomical Record.Part A, Discoveries in Molecular, Cellular, and Evolutionary Biology* 283 (1) (Mar): 187-92.

Leahy, A., J. W. Xiong, F. Kuhnert, and H. Stuhlmann. 1999. Use of developmental marker genes to define temporal and spatial patterns of differentiation during embryoid body formation. *The Journal of Experimental Zoology* 284 (1) (Jun 15): 67-81.

Lee, Ji-Yeon, Hyehyun Min, Xinnan Wang, Abdul Aziz Khan, and Myoung Hee Kim. 2010. Chromatin organization and transcriptional activation of hox genes. *Anat Cell Biol* 43 (1) (/3/): 78-85.

Lempradl, A., and L. Ringrose. 2008. How does noncoding transcription regulate hox genes? *BioEssays : News and Reviews in Molecular, Cellular and Developmental Biology* 30 (2) (Feb): 110-21.

LEWIS, E. B. 1951. Pseudoallelism and gene evolution. *Cold Spring Harbor Symposia on Quantitative Biology* 16 : 159-74.

LEWIS, E. B. 1954. The theory and application of a new method of detecting chromosomal rearrangements in Drosophila melanogaster. T*he American naturalist,* Vol. 88, No. 841: 225-239

LEWIS, E. B. 1963. Genes and Developmental Pathways. *American Zoologist.* Vol. 3, No. 1 , 33-56

Lin, C., E. R. Smith, H. Takahashi, K. C. Lai, S. Martin-Brown, L. Florens, M. P. Washburn, J. W. Conaway, R. C. Conaway, and A. Shilatifard. 2010. AFF4, a component of the ELL/P-TEFb elongation complex and a shared subunit of MLL chimeras, can link transcription elongation to leukemia. *Molecular Cell* 37 (3) (Feb 12): 429-37.

Ling, V., and S. Neben. 1997. In vitro differentiation of embryonic stem cells: Immunophenotypic analysis of cultured embryoid bodies. *Journal of Cellular Physiology* 171 (1) (Apr): 104-15.

Lipshitz, H. D., D. A. Peattie, and D. S. Hogness. 1987. Novel transcripts from the ultrabithorax domain of the bithorax complex. *Genes & Development* 1 (3) (May): 307-22.

Loh, Y. H., Q. Wu, J. L. Chew, V. B. Vega, W. Zhang, X. Chen, G. Bourque, et al. 2006. The Oct4 and nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nature Genetics* 38 (4) (Apr): 431-40.

Lu, Shi-Jiang, Chengshi Quan, Fei Li, Loyda Vida, and George R. Honig. 2002. Hematopoietic progenitor cells derived from embryonic stem cells: Analysis of gene expression. *Stem Cells* 20 (5): 428-37.

Maeda, R. K., and F. Karch. 2006. The ABC of the BX-C: The bithorax complex explained. *Development (Cambridge, England)* 133 (8) (Apr): 1413-22.

Magli, M. C., C. Largman, and H. J. Lawrence. 1997. Effects of HOX homeobox genes in blood cell differentiation. *Journal of Cellular Physiology* 173 (2) (Nov): 168-77.

Maris, C., C. Dominguez, and F. H. Allain. 2005. The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. *FEBS Journal* 272 (9): 2118-31.

Marschalek, R. 2011. Mechanisms of leukemogenesis by MLL fusion proteins. *British Journal of Haematology* 152 (2) (Jan): 141-54.

Matharu, N. K., T. Hussain, R. Sankaranarayanan, and R. K. Mishra. 2010. Vertebrate homologue of drosophila GAGA factor. *Journal of Molecular Biology* 400 (3) (Jul 16): 434-47.

Mattick, J. S. 2009. The genetic signatures of noncoding RNAs. *PLoS Genetics* 5 (4) (Apr): e1000459.

———. 2001. Non-coding RNAs: The architects of eukaryotic complexity. *EMBO Reports* 2 (11) (Nov): 986-91.

Matys, V., E. Fricke, R. Geffers, E. Gossling, M. Haubrock, R. Hehl, K. Hornischer, et al. 2003. TRANSFAC: Transcriptional regulation, from patterns to profiles. *Nucleic Acids Research* 31 (1) (Jan 1): 374-8.

McCabe, N. R., R. C. Burnett, H. J. Gill, M. J. Thirman, D. Mbangkollo, M. Kipiniak, E. van Melle, S. Ziemin-van der Poel, J. D. Rowley, and M. O. Diaz. 1992. Cloning of cDNAs of the MLL gene that detect DNA rearrangements and altered RNA transcripts in human leukemic cells with 11q23 translocations. *Proceedings of the National Academy of Sciences of the United States of America* 89 (24): 11794-8.

Mendenhall, E. M., and B. E. Bernstein. 2008. Chromatin state maps: New technologies, new insights. *Current Opinion in Genetics & Development* 18 (2) (Apr): 109-15.

Mercer, T. R., M. E. Dinger, S. M. Sunkin, M. F. Mehler, and J. S. Mattick. 2008. Specific expression of long noncoding RNAs in the mouse brain. *Proceedings of the National Academy of Sciences of the United States of America* 105 (2) (Jan 15): 716-21.

Mi, H., O. Kops, E. Zimmermann, A. Jaschke, and M. Tropschug. 1996. A nuclear RNA-binding cyclophilin in human T cells. *FEBS Letters* 398 (2-3): 201-5.

Mihaly, J., I. Hogga, S. Barges, M. Galloni, R. K. Mishra, K. Hagstrom, M. Muller, et al. 1998. Chromatin domain boundaries in the bithorax complex. *Cellular and Molecular Life Sciences : CMLS* 54 (1) (Jan): 60-70.

Milne, T. A., S. D. Briggs, H. W. Brock, M. E. Martin, D. Gibbs, C. D. Allis, and J. L. Hess. 2002. MLL targets SET domain methyltransferase activity to hox gene promoters. *Molecular Cell* 10 (5) (Nov): 1107-17.

Milne, T. A., Y. Dou, M. E. Martin, H. W. Brock, R. G. Roeder, and J. L. Hess. 2005. MLL associates specifically with a subset of transcriptionally active target genes. *Proceedings of the National Academy of Sciences of the United States of America* 102 (41) (Oct 11): 14765-70.

Mohan, M., C. Lin, E. Guest, and A. Shilatifard. 2010. Licensed to elongate: A molecular mechanism for MLL-based leukaemogenesis. *Nature Reviews.Cancer* 10 (10) (Oct): 721-8.

Muntean, A. G., D. Giannola, A. M. Udager, and J. L. Hess. 2008. The PHD fingers of MLL block MLL fusion protein-mediated transformation. *Blood* 112 (12) (Dec 1): 4690-3.

Muntean, A. G., J. Tan, K. Sitwala, Y. Huang, J. Bronstein, J. A. Connelly, V. Basrur, K. S. Elenitoba-Johnson, and J. L. Hess. 2010. The PAF complex synergizes with MLL fusion proteins at HOX loci to promote leukemogenesis. *Cancer Cell* 17 (6) (Jun 15): 609-21.

Nakamura, T., T. Mori, S. Tada, W. Krajewski, T. Rozovskaia, R. Wassell, G. Dubois, A. Mazo, C. M. Croce, and E. Canaani. 2002. ALL-1 is a histone methyltransferase that assembles a supercomplex of proteins involved in transcriptional regulation. *Molecular Cell* 10 (5) (Nov): 1119-28.

Nielsen, Rasmus, Adam Siepel, and David Haussler. 2005. Phylogenetic hidden markov models. In *Statistical methods in molecular evolution.*, 325-351; 351Springer New York.

Nolis, I. K., D. J. McKay, E. Mantouvalou, S. Lomvardas, M. Merika, and D. Thanos. 2009. Transcription factors mediate long-range enhancer-promoter interactions. *Proceedings of the National Academy of Sciences of the United States of America* 106 (48) (Dec 1): 20222-7.

Oliver, G., C. V. Wright, J. Hardwicke, and E. M. De Robertis. 1988. Differential antero-posterior expression of two proteins encoded by a homeobox gene in xenopus and mouse embryos. *The EMBO Journal* 7 (10) (Oct): 3199-209.

Orlando, V., E. P. Jane, V. Chinwalla, P. J. Harte, and R. Paro. 1998. Binding of trithorax and polycomb proteins to the bithorax complex: Dynamic changes during early drosophila embryogenesis. *The EMBO Journal* 17 (17) (Sep 1): 5141-50.

Park, S., U. Osmers, G. Raman, R. H. Schwantes, M. O. Diaz, and J. H. Bushweller. 2010. The PHD3 domain of MLL acts as a CYP33-regulated switch between MLL-mediated activation and repression. *Biochemistry* 49 (31) (Aug 10): 6576-86.

Petruk, S., Y. Sedkov, H. W. Brock, and A. Mazo. 2007. A model for initiation of mosaic HOX gene expression patterns by non-coding RNAs in early embryos. *RNA Biology* 4 (1) (Jan-Mar): 1-6.

Petruk, S., Y. Sedkov, K. M. Riley, J. Hodgson, F. Schweisguth, S. Hirose, J. B. Jaynes, H. W. Brock, and A. Mazo. 2006. Transcription of bxd noncoding RNAs promoted by trithorax represses ubx in cis by transcriptional interference. *Cell* 127 (6) (Dec 15): 1209-21.

Plant, K. E., S. J. Routledge, and N. J. Proudfoot. 2001. Intergenic transcription in the human beta-globin gene cluster. *Molecular and Cellular Biology* 21 (19) (Oct): 6507-14.

Rada-Iglesias, A., R. Bajpai, T. Swigut, S. A. Brugmann, R. A. Flynn, and J. Wysocka. 2011. A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* 470 (7333) (Feb 10): 279-83.

Rank, G., M. Prestel, and R. Paro. 2002. Transcription through intergenic chromosomal memory elements of the drosophila bithorax complex correlates with an epigenetic switch. *Molecular and Cellular Biology* 22 (22) (Nov): 8026-34.

Rea, S., G. Xouri, and A. Akhtar. 2007. Males absent on the first (MOF): From flies to humans. *Oncogene* 26 (37) (Aug 13): 5385-94.

Ringrose, L., and R. Paro. 2007. Polycomb/Trithorax response elements and epigenetic memory of cell identity. *Development (Cambridge, England)* 134 (2) (Jan): 223-32.

Rinn, J. L., M. Kertesz, J. K. Wang, S. L. Squazzo, X. Xu, S. A. Brugmann, L. H. Goodnough, et al. 2007. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* 129 (7) (Jun 29): 1311-23.

Rogan, D. F., D. J. Cousins, and D. Z. Staynov. 1999. Intergenic transcription occurs throughout the human IL-4/IL-13 gene cluster. *Biochemical and Biophysical Research Communications* 255 (3) (Feb 24): 556-61.

Ronshaugen, M., and M. Levine. 2004. Visualization of trans-homolog enhancer-promoter interactions at the abd-B hox locus in the drosophila embryo. *Developmental Cell* 7 (6) (Dec): 925-32.

Rozenblatt-Rosen, O., T. Rozovskaia, D. Burakov, Y. Sedkov, S. Tillib, J. Blechman, T. Nakamura, C. M. Croce, A. Mazo, and E. Canaani. 1998. The C-terminal SET domains of ALL-1 and TRITHORAX interact with the INI1 and SNR1 proteins, components of the SWI/SNF complex. *Proceedings of the National Academy of Sciences of the United States of America* 95 (8) (Apr 14): 4152-7.

Sabarinadh, C., S. Subramanian, A. Tripathi, and R. K. Mishra. 2004. Extreme conservation of noncoding DNA near HoxD complex of vertebrates. *BMC Genomics* 5 (Oct 6): 75.

Sakashita, E., and H. Sakamoto. 1994. Characterization of RNA binding specificity of the drosophila sex-lethal protein by in vitro ligand selection. *Nucleic Acids Research* 22 (20) (Oct 11): 4082-6.

Sanchez-Elsner, T., D. Gou, E. Kremmer, and F. Sauer. 2006. Noncoding RNAs of trithorax response elements recruit drosophila Ash1 to ultrabithorax. *Science (New York, N.Y.)* 311 (5764) (Feb 24): 1118-23.

Schmitt, R. M., E. Bruyns, and H. R. Snodgrass. 1991. Hematopoietic development of embryonic stem cells in vitro: Cytokine and receptor gene expression. *Genes & Development* 5 (5) (May): 728-40.

Schuettengruber, B., D. Chourrout, M. Vervoort, B. Leblanc, and G. Cavalli. 2007. Genome regulation by polycomb and trithorax proteins. *Cell* 128 (4) (Feb 23): 735-45.

Schuettengruber, B., M. Ganapathi, B. Leblanc, M. Portoso, R. Jaschek, B. Tolhuis, M. van Lohuizen, A. Tanay, and G. Cavalli. 2009. Functional anatomy of polycomb and trithorax chromatin landscapes in drosophila embryos. *PLoS Biology* 7 (1) (Jan 13): e13.

Schwartz, Y. B., T. G. Kahn, P. Stenberg, K. Ohno, R. Bourgon, and V. Pirrotta. 2010. Alternative epigenetic chromatin states of polycomb target genes. *PLoS Genetics* 6 (1) (Jan): e1000805.

Scott, M. P. 1987. Complex loci of drosophila. *Annual Review of Biochemistry* 56 (1) (06/01; 2011/08): 195-227.

Shah, N., and S. Sukumar. 2010. The hox genes and their roles in oncogenesis. *Nature Reviews.Cancer* 10 (5) (May): 361-71.

Shashikant, C. S., C. J. Bieberich, H. G. Belting, J. C. Wang, M. A. Borbely, and F. H. Ruddle. 1995. Regulation of hoxc-8 during mouse embryonic development: Identification and characterization of critical elements involved in early neural tube expression. *Development (Cambridge, England)* 121 (12) (Dec): 4339-47.

Shashikant, C. S., and F. H. Ruddle. 1996. Combinations of closely situated cis-acting elements determine tissue-specific patterns and anterior extent of early Hoxc8 expression. *Proceedings of the National Academy of Sciences of the United States of America* 93 (22) (Oct 29): 12364-9.

Shimamoto, T., K. Ohyashiki, K. Toyama, and K. Takeshita. 1998. Homeobox genes in hematopoiesis and leukemogenesis. *International Journal of Hematology* 67 (4) (Jun): 339-50.

Shimamoto, T., Y. Tang, Y. Naot, M. Nardi, P. Brulet, C. J. Bieberich, and K. Takeshita. 1999. Hematopoietic progenitor cell abnormalities in hoxc-8 null mutant mice. *The Journal of Experimental Zoology* 283 (2) (Feb 1): 186-93.

Slany, R. K. 2009. The molecular biology of mixed lineage leukemia. *Haematologica* 94 (7) (Jul): 984-93.

Smith, S. T., S. Petruk, Y. Sedkov, E. Cho, S. Tillib, E. Canaani, and A. Mazo. 2004. Modulation of heat shock gene expression by the TAC1 chromatin-modifying complex. *Nature Cell Biology* 6 (2) (Feb): 162-7.

Takahashi, Yoko, Jun-ichi Hamada, Katsuhiko Murakawa, Minoru Takada, Mitsuhiro Tada, Ikuko Nogami, Nobuyasu Hayashi, et al. 2004. Expression profiles of 39 HOX genes in normal human adult organs and anaplastic thyroid cancer cell lines by quantitative real-time RT-PCR system. *Experimental Cell Research* 293 (1) (/2/1/): 144-53.

Tolhuis, B., M. Blom, R. M. Kerkhoven, L. Pagie, H. Teunissen, M. Nieuwland, M. Simonis, W. de Laat, M. van Lohuizen, and B. van Steensel. 2011. Interactions among polycomb domains are guided by chromosome architecture. *PLoS Genetics* 7 (3) (Mar): e1001343.

Tsai, M. C., O. Manor, Y. Wan, N. Mosammaparast, J. K. Wang, F. Lan, Y. Shi, E. Segal, and H. Y. Chang. 2010. Long noncoding RNA as modular scaffold of histone modification complexes. *Science (New York, N.Y.)* 329 (5992) (Aug 6): 689-93.

Tschopp, P., B. Tarchini, F. Spitz, J. Zakany, and D. Duboule. 2009. Uncoupling time and space in the collinear regulation of hox genes. *PLoS Genetics* 5 (3) (Mar): e1000398.

Tupy, J. L., A. M. Bailey, G. Dailey, M. Evans-Holm, C. W. Siebel, S. Misra, S. E. Celniker, and G. M. Rubin. 2005. Identification of putative noncoding polyadenylated transcripts in drosophila melanogaster. *Proceedings of the National Academy of Sciences of the United States of America* 102 (15) (Apr 12): 5495-500.

Venter, J. C., Mark D. Adams, Eugene W. Myers, Peter W. Li, Richard J. Mural, Granger G. Sutton, Hamilton O. Smith, et al. 2001. The sequence of the human genome. *Science* 291 (5507) (02/16): 1304-51.

Wang, K. C., Y. W. Yang, B. Liu, A. Sanyal, R. Corces-Zimmerman, Y. Chen, B. R. Lajoie, et al. 2011. A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature* 472 (7341) (Apr 7): 120-4.

Wang, P., C. Lin, E. R. Smith, H. Guo, B. W. Sanderson, M. Wu, M. Gogol, et al. 2009. Global analysis of H3K4 methylation defines MLL family member targets and points to a role for MLL1-mediated H3K4 methylation in the regulation of transcriptional initiation by RNA polymerase II. *Molecular & Cellular Biology* 29 (22): 6074-85.

Wang, Y., R. Han, W. Zhang, Y. Yuan, X. Zhang, Y. Long, and H. Mi. 2008. Human CyP33 binds specifically to mRNA and binding stimulates PPIase activity of hCyP33. *FEBS Letters* 582 (5): 835-9.

Wang, Z., J. Song, T. A. Milne, G. G. Wang, H. Li, C. D. Allis, and D. J. Patel. 2010. Pro isomerization in MLL1 PHD3-bromo cassette connects H3K4me readout to CyP33 and HDAC-mediated repression. *Cell* 141 (7) (Jun 25): 1183-94.

Wobus, A. M., and K. R. Boheler. 2005. Embryonic stem cells: Prospects for developmental biology and cell therapy. *Physiological Reviews* 85 (2) (Apr): 635-78.

Wu, D. Y., D. C. Tkachuck, R. S. Roberson, and W. H. Schubach. 2002. The human SNF5/INI1 protein facilitates the function of the growth arrest and DNA damage-inducible protein (GADD34) and modulates GADD34-bound protein phosphatase-1 activity. *The Journal of Biological Chemistry* 277 (31) (Aug 2): 27706-15.

Xia, Z. B., M. Anderson, M. O. Diaz, and N. J. Zeleznik-Le. 2003. MLL repression domain interacts with histone deacetylases, the polycomb group proteins HPC2 and BMI-1, and the corepressor C-terminal-binding protein. *Proceedings of the National Academy of Sciences of the United States of America* 100 (14) (Jul 8): 8342-7.

Xiang, P., X. Fang, W. Yin, G. Barkess, and Q. Li. 2006. Non-coding transcripts far upstream of the epsilon-globin gene are distinctly expressed in human primary tissues and erythroleukemia cell lines. *Biochemical and Biophysical Research Communications* 344 (2) (Jun 2): 623-30.

Yano, T., T. Nakamura, J. Blechman, C. Sorio, C. V. Dang, B. Geiger, and E. Canaani. 1997. Nuclear punctate distribution of ALL-1 is conferred by distinct elements at the N terminus of the protein. *Proceedings of the National Academy of Sciences of the United States of America* 94 (14): 7286-91.

Yokoyama, A., and M. L. Cleary. 2008. Menin critically links MLL proteins with LEDGF on cancer-associated target genes. *Cancer Cell* 14 (1) (Jul 8): 36-46.

Yokoyama, A., M. Lin, A. Naresh, I. Kitabayashi, and M. L. Cleary. 2010 A higher-order complex containing AF4 and ENL family proteins with P-TEFb facilitates oncogenic and physiologic MLL-dependent transcription. *Cancer Cell* 17 (2): 198-212.

Yokoyama, A., T. C. Somervaille, K. S. Smith, O. Rozenblatt-Rosen, M. Meyerson, and M. L. Cleary. 2005. The menin tumor suppressor protein is an essential oncogenic cofactor for MLL-associated leukemogenesis. *Cell* 123 (2) (Oct 21): 207-18.

Yokoyama, A., Z. Wang, J. Wysocka, M. Sanyal, D. J. Aufiero, I. Kitabayashi, W. Herr, and M. L. Cleary. 2004. Leukemia proto-oncoprotein MLL forms a SET1-like histone methyltransferase complex with menin to regulate hox gene expression. *Molecular and Cellular Biology* 24 (13) (Jul): 5639-49.

Yu, B. D., R. D. Hanson, J. L. Hess, S. E. Horning, and S. J. Korsmeyer. 1998. MLL, a mammalian trithorax-group gene, functions as a transcriptional maintenance factor in morphogenesis. *Proceedings of the National Academy of Sciences of the United States of America* 95 (18) (Sep 1): 10632-6.

Yu, B. D., J. L. Hess, S. E. Horning, G. A. Brown, and S. J. Korsmeyer. 1995. Altered hox expression and segmental identity in mll-mutant mice. *Nature* 378 (6556) (Nov 30): 505-8.

Zeleznik-Le, N. J., A. M. Harden, and J. D. Rowley. 1994. 11q23 translocations split the "AT-hook" cruciform DNA-binding region and the transcriptional repression domain from the activation domain of the mixed-lineage leukemia (MLL) gene. *Proceedings of the National Academy of Sciences of the United States of America* 91 (22) (Oct 25): 10610-4.

Zhang, X., Z. Lian, C. Padden, M. B. Gerstein, J. Rozowsky, M. Snyder, T. R. Gingeras, P. Kapranov, S. M. Weissman, and P. E. Newburger. 2009. A myelopoiesis-associated regulatory intergenic noncoding RNA transcript within the human HOXA cluster. *Blood* 113 (11) (Mar 12): 2526-34.

Zhang, Y., and J. D. Rowley. 2006. Chromatin structural elements and chromosomal translocations in leukemia. *DNA Repair* 5 (9-10): 1282-97.

Zhou, J., H. Ashe, C. Burks, and M. Levine. 1999. Characterization of the transvection mediating region of the abdominal-B locus in drosophila. *Development (Cambridge, England)* 126 (14) (Jun): 3057-65.

Zhou, J., S. Barolo, P. Szymanski, and M. Levine. 1996. The fab-7 element of the bithorax complex attenuates enhancer-promoter interactions in the drosophila embryo. *Genes & Development* 10 (24) (Dec 15): 3195-201.

Ziemin-van der Poel, S., N. R. McCabe, H. J. Gill, R. Espinosa 3rd, Y. Patel, A. Harden, P. Rubinelli, S. D. Smith, M. M. LeBeau, and J. D. Rowley. 1991. Identification of a gene, MLL, that spans the breakpoint in 11q23 translocations associated with human leukemias. *Proceedings of the National Academy of Sciences of the United States of America* 88 (23) (Dec 1): 10735-9.

Zuker, M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research* 31 (13): 3406-15.

VITA

Jessica Solanki, was born on August 26, 1977, in Gujarat, India to Arvind and Ranjan Solanki. Jessica obtained her Bachelor's degree in Microbiology (Honors) from the University of Mumbai, India. In order to gain a better understanding of Life Sciences, she enrolled at the University of Pune in the Biotechnology program which allowed for a more comprehensive curriculum. During her studies in Biotechnology, she got an opportunity to participate in a one year agriculture based project in the Department of Plant Biotechnology, National Chemical Laboratory, Pune. Jessica received hands-on training and an understanding of plant genetics based on identification of quantitative trait-specific DNA markers. She graduated with a Master's in Biotechnology in 2000.

In August 2001, Jessica enrolled at the Department of Cell Biology, Neurobiology and Anatomy at Loyola University Chicago for the Doctorate in Philosophy degree. In the January of 2002, she changed her PhD major to Molecular and Cellular Biochemistry. With an interest in cancer biology, she joined the laboratory of Dr. Manuel O. Diaz (M.D). She accepted a project in the lab that addressed whether or not the non-coding RNA, known to be transcribed by spacer DNA in organisms ranging from protists to humans, has a role in gene regulation by the MLL protein. During the course of her graduate studies, Jessica was a recipient of the prestigious Arthur J. Schmitt award

during the academic year 2005-2006. She also served as a Secretary of the Graduate Student Council at Loyola University Medical Center during the academic year 2004-2005.